Mass appeal

Dunn, Warwick B.; Erban, Alexander; Weber, Ralf J.M.; Creek, Darren J.; Brown, Marie; Breitling, Rainer; Hankemeier, Thomas; Goodacre, Royston; Neumann, Steffen; Kopka, Joachim

Published in:
Metabolomics

DOI:
10.1007/s11306-012-0434-4

Publication date:
2013

Metabolomics (2013) 9:S44–S66
DOI 10.1007/s11306-012-0434-4

ORIGINAL ARTICLE

Mass appeal: metabolite identification in mass spectrometry-focused untargeted metabolomics

Warwick B. Dunn · Alexander Erban · Ralf J. M. Weber · Darren J. Creek · Marie Brown · Rainer Breitling · Thomas Hankemeier · Royston Goodacre · Steffen Neumann · Joachim Kopka · Mark R. Viant

Received: 23 March 2012 / Accepted: 10 May 2012 / Published online: 26 May 2012

Abstract Metabolomics has advanced significantly in the past 10 years with important developments related to hardware, software and methodologies and an increasing complexity of applications. In discovery-based investigations, applying untargeted analytical methods, thousands of metabolites can be detected with no or limited prior knowledge of the metabolite composition of samples. In these cases, metabolite identification is required following data acquisition and processing. Currently, the process of metabolite identification in untargeted metabolomic studies is a significant bottleneck in deriving biological knowledge from metabolomic studies. In this review we highlight the different traditional and emerging tools and strategies applied to identify subsets of metabolites detected in untargeted metabolomic studies applying various mass spectrometry platforms. We indicate the workflows which are routinely applied and highlight the current limitations which need to be overcome to provide efficient, accurate and robust identification of metabolites in untargeted metabolomic studies. These workflows apply to the identification of metabolites, for which the structure can be assigned based on entries in databases, and for those which

W. B. Dunn (✉) · M. Brown
Centre for Advanced Discovery & Experimental Therapeutics (CADET), Central Manchester NHS Foundation Trust, Manchester Academic Health Sciences Centre, University of Manchester, York Place, Oxford Road, Manchester M13 9WL, UK
e-mail: warwick.dunn@manchester.ac.uk

W. B. Dunn · M. Brown
School of Biomedicine, University of Manchester, Oxford Road, Manchester M13 9PL, UK

A. Erban · J. Kopka
Max Planck Institute for Molecular Plant Physiology (MPIMP), Potsdam-Golm, Germany

R. J. M. Weber · M. R. Viant
Centre for Systems Biology, University of Birmingham, Edgbaston, Birmingham B15 2TT, UK

R. J. M. Weber · M. R. Viant
School of Biosciences, University of Birmingham, Edgbaston, Birmingham B15 2TT, UK

D. J. Creek
Institute of Infection, Immunity and Inflammation, College of Medical, Veterinary and Life Sciences, University of Glasgow, Glasgow, UK

D. J. Creek
Department of Biochemistry and Molecular Biology, University of Melbourne, Parkville, VIC, Australia

R. Breitling
Institute of Molecular, Cell and Systems Biology, College of Medical, Veterinary and Life Sciences, University of Glasgow, Glasgow, UK

R. Breitling
Groningen Bioinformatics Centre, Groningen Biomolecular Sciences and Biotechnology Institute, University of Groningen, Groningen, The Netherlands

T. Hankemeier
Division of Analytical Biosciences, LACDR, Leiden University, P.O. Box 9502, 2300 RA Leiden, The Netherlands

T. Hankemeier
Netherlands Metabolomics Centre, LACDR, Leiden University, P.O. Box 9502, 2300 RA Leiden, The Netherlands

R. Goodacre
Manchester Centre for Integrative Systems Biology, University of Manchester, 131 Princess Street, Manchester M1 7DN, UK

are not yet stored in databases and which require a de novo structure elucidation.

Keywords Capillary electrophoresis · Metabolomics · Metabolite identification · Structure elucidation · Mass spectrometry · Gas chromatography · Liquid chromatography · Ultra performance liquid chromatography · DIMS

1 Introduction

At the turn of the twenty-first century, advances in analytical and informatics technologies provided the drive for the emerging scientific field of metabolomics to develop and rapidly grow. Today, metabolomics tools are applied in the investigation of microbial (Pope et al. 2007; Yuan et al. 2009; Kahar et al. 2011; Winder et al. 2011), plant (Roessner et al. 2001; Farag et al. 2009; Lugan et al. 2010), environmental (Viant 2008; Boroujerdi et al. 2009) and mammalian (Oresic et al. 2008; Dunn et al. 2011a) systems. Metabolomics in these systems has a diverse range of scientific objectives including the discovery of biomarkers through to the understanding of biological mechanisms related to genetic and/or environmental perturbations.

The experimental strategies applied in these studies can be categorized into three classes: targeted analysis, semi-targeted analysis and untargeted analysis (also known as metabolic profiling, metabolite profiling or metabolomics) (Fiehn 2002; Dunn et al. 2011a). These strategies differ in many aspects including the level of quantitation (relative vs. absolute), complexity of sample preparation, experimental accuracy and precision, number of metabolites detected, and the study objective (hypothesis generation/discovery study vs. hypothesis testing study). One major difference when comparing untargeted analysis to targeted and semi-targeted analyses is the need for chemical identification and structural elucidation of detected metabolites. For targeted and semi-targeted analyses, the chemical identities of the metabolite or metabolites to be assayed are known before data acquisition commences, and analytical methods are developed to provide high accuracy, precision and selectivity. These methods are developed with the application of authentic chemical standards. The subsequent process of deriving biological knowledge from acquired data can be started immediately following data analysis as the chemical identity of the metabolites is known. This is a significant advantage of these strategies, though fewer metabolites are typically detected and reported, compared to untargeted analysis, and this may not be appropriate for true discovery studies.

In untargeted analyses, fit-for-purpose analytical methods are developed to acquire data on a diverse range of metabolites. Specific knowledge of which metabolites will be detected prior to data acquisition is limited, though information related to *metabolite classes* of interest can allow an appropriate choice of analytical platform and sample preparation method to provide enrichment of metabolites of interest. The chemical identification and structural elucidation of all, or more realistically, many biologically interesting metabolites, is a labour-intensive step that follows data acquisition and analysis and must occur before biological interpretation is possible. It is, therefore, important to realize that metabolite identification in untargeted analysis has been highlighted repeatedly as a significant bottleneck in mass spectrometry-focused metabolomic studies (Dunn et al. 2011a; Wishart 2011); a survey at *The American Society for Mass Spectrometry* annual conference in 2009 provided evidence of this bottleneck across a wide set of researchers (http://fiehnlab.ucdavis.edu/staff/kind/Metabolomics-Survey-2009/).

Untargeted metabolomics studies typically apply mass spectrometry coupled to a range of diverse chromatographic platforms, including gas chromatography (GC–MS and comprehensive GC × GC; (Fiehn et al. 2000a; Roessner et al. 2001; Welthagen et al. 2005; Huege et al. 2011), liquid chromatography (LC–MS), and related advanced hardware including ultra-performance liquid chromatography (UPLC–MS, also referred to as ultra-high performance liquid chromatography—UHPLC–MS; Theodoridis et al. 2008; Brown et al. 2009; Dunn et al. 2009; Spagou et al. 2010), and capillary electrophoresis (CE–MS; Soga et al. 2003; Ramautar et al. 2009). Alternatively, samples can be directly injected or infused into the mass spectrometer (direct infusion/injection mass spectrometry, DIMS; Southam et al. 2007; Beckmann et al. 2008; Taylor et al. 2009; Fuhrer et al. 2011; Weber et al. 2011). The advantages and limitations of the different mass spectrometry platforms have been extensively reviewed (Dettmer et al. 2007; Dunn 2008; Lei et al. 2011). Despite recent technological advancements in all mass spectrometry platforms, no single analytical platform or manufacturer's instrument is the perfect tool for untargeted metabolomics, all having advantages and limitations.

When applied to pure chemicals or to relatively simple mixtures, mass spectrometry offers a range of powerful tools that can be used for the characterization, structural

R. Goodacre
School of Chemistry, Manchester Interdisciplinary Biocentre, University of Manchester, 131 Princess Street, Manchester M1 7DN, UK

S. Neumann
Department of Stress and Developmental Biology, Leibniz Institute of Plant Biochemistry, 06120 Halle, Germany

elucidation, and identification of metabolites. These include (i) the accurate measurement of the mass-to-charge ratio (*m/z*) of molecular, fragment and associated ions; (ii) the determination of relative isotopic abundances (RIAs) (e.g., the relative abundance of ^{12}C and ^{13}C isotopomers) of molecular and fragment ions; (iii) fragmentation of molecular and fragment ions to define dissociation patterns related to chemical structure; and (iv) the comparison of experimental data to either databases containing physicochemical properties (e.g., molecular formulas and monoisotopic masses) or mass spectral libraries containing experimentally acquired chromatographic [e.g., retention times (RTs) or retention indices] and mass spectrometry data (e.g., fragmentation mass spectra). As an example, Fiehn and Kind have defined how the "seven golden rules" of traditional analytical chemistry can be applied in metabolite identification (Kind and Fiehn 2007). Data applied for the identification or annotation of metabolites can be collected in two different processes; (a) during the data acquisition step of untargeted metabolomics (for example, RT and electron-impact mass spectrum data for GC–MS and *m/z*, MS/MS mass spectrum and RT data for LC–MS) or (b) can be collected in a targeted manner following the data acquisition, processing and analysis stages (for example, acquisition of MS^n data for LC eluent fractions collected during the data acquisition stage of untargeted metabolomics).

However, the challenge of metabolite identification is still considerable in untargeted metabolomic studies. Samples are complex and can contain hundreds or thousands of chemical species, depending on the biological system and sample type being studied. For example, biofluids acquired from the human population contain endogenous metabolites as well as exogenous metabolites derived from diet (Lloyd et al. 2011), lifestyle and physical activity (Pechlivanis et al. 2010), pharmaceuticals (Loo et al. 2012), and the gut microflora (Wikoff et al. 2009), most of them at low concentrations (micromolar or lower). Complex mammalian systems are affected by many intrinsic and extrinsic factors and can be thought of as superorganisms (Goodacre 2007). From a knowledge perspective, the *total* qualitative composition of many metabolomes is currently incomplete. Moreover, it is often not known which metabolites should be present in a sample; databases are available which contain large lists of the expected metabolites in different organisms, based on experimental, genomic and/or bibliographic data (for example, The Human Metabolome Database—HMDB; Wishart et al. 2009) and the metabolic reconstruction of yeast (Herrgård et al. 2008), but these lists are far from complete. Finally, the physicochemical diversity of the metabolome is significantly greater than that of the proteome (Wishart 2011), making generally applicable identification strategies all but impossible. Characterization of peptides and proteins, which are linear polymers composed of about 20 amino acids, is significantly simpler than characterizing the complex structural arrangements observed in metabolites; although when modified post-translationally (e.g., phosphate, glycans, etc.) protein identification is challenging. All the points defined above provide difficulties in chemically characterizing all detected metabolites or more realistically a subset of biologically interesting metabolites in a semi-automated or automated process.

2 Challenges and requirements of metabolite identification

The identification of metabolites in metabolomic samples has to discriminate (i) metabolites of different nominal mass; (ii) metabolites with the same nominal mass but different molecular formula and monoisotopic mass; and (iii) metabolites with the same nominal and monoisotopic masses, but different chemical structures (including chirality and isomerism; for example, leucine and isoleucine are isomers with the same nominal and monoisotopic masses). Furthermore, as single metabolites are usually detected in a mass spectrometer as multiple different derived species, correct assignment to the "parent" metabolite is essential. For example, in GC–MS, chemical derivatisation by trimethylsilylation (TMS) reagents can result in detection of amino acids containing 1, 2 or 3 TMS groups; (Halket and Zaikin 2003), and in data acquired from electrospray ionization (ESI) mass spectrometers a single metabolite can form multiple different ion types (e.g., sodium and potassium adduct ions, in addition to the standard protonated form) (Brown et al. 2009). Each different detected form of a metabolite is commonly referred to as a metabolic feature or a metabolite feature.

The identification challenge is, therefore, immense, and confident unambiguous assignments of observed metabolic features to a single metabolite are not always achievable. The Chemical Analysis Working Group of the Metabolomics Standards Initiative (MSI; http://msi-workgroups.sourceforge.net) has defined four different levels of metabolite identification confidence (as defined in Table 1) and methods on how to report metabolite identities (Sumner et al. 2007).

Definitive (level 1) identification requires comparison of two or more orthogonal properties (e.g., RT/index, *m/z*, fragmentation mass spectrum) of an authentic chemical standard to the same properties observed for the metabolite of interest analysed under identical analytical conditions (in the researcher's laboratory or a separate laboratory). The probability of an accurate identification is high, but

Table 1 The four levels of metabolite identification confidence defined by the Metabolomics Standards Initiative (Sumner et al. 2007)

Level	Confidence of identity	Level of evidence
1	Confidently identified compounds	Comparison of two or more orthogonal properties with an authentic chemical standard analysed under identical analytical conditions
2	Putatively annotated compounds	Based upon physicochemical properties and/or spectral similarity with public/commercial spectral libraries, without reference to authentic chemical standards
3	Putatively annotated compound classes	Based upon characteristic physicochemical properties of a chemical class of compounds, or by spectral similarity to known compounds of a chemical class
4	Unknown compounds	Although unidentified and unclassified, these metabolites can still be differentiated and quantified based upon spectral data

even if this level of information is available it may be impossible to distinguish some metabolites. Many isomers, especially stereoisomers, appear very similar or identical based on chromatographic or mass spectrometric characteristics, particularly in non-optimized and rapid analysis methods that are commonly applied in untargeted metabolomic studies. If an accurate identification of isomers is required, and/or the presence of a mixture is suspected, the development of chromatographic methods that unambiguously separate different stereoisomers is required. Here, NMR spectroscopy can be very powerful in determining structural configurations of stereoisomers.

Putative (level 2 or 3) annotation is typically based on one or two properties only and often relies on comparison to data collected in different laboratories and acquired with different analytical methods, instead of a direct comparison with an authentic chemical standard under identical analytical conditions. The properties used depend on the platform: e.g., in GC–MS the electron impact (EI) fragmentation mass spectra contained in GC–MS mass spectral libraries can be applied for putative annotation; the fragmentation patterns are directly comparable because the mechanism of fragmentation is reproducible across many different GC–MS platforms. In LC–MS and UPLC–MS, accurately measured *m/z* is typically the first property used to identify metabolites, and may be combined with comparison of fragmentation spectra or RTs against experimentally or computationally derived databases. Again, this is not based on a comparison to an authentic chemical standard applying an identical analytical method, and therefore the resulting identifications are defined as putative.

The difficulties in metabolite identification discussed above and the current lack of full qualitative descriptions of sample-specific metabolomes leads to two different types of identification strategies being applied, (i) assignment of the identity of a metabolite based on data stored in databases, and (ii) de novo structure elucidation where no data on the metabolite can be found in any database.

3 Metabolite identification in GC–MS and comprehensive GC × GC–MS derived datasets

3.1 Application of EI mass spectra and retention indices

The majority of GC–MS and GC × GC–MS platforms applied in metabolomic studies operate with EI ionization. The metabolite identification workflow in untargeted metabolomic studies typically applies the RT/retention index and/or the EI-derived mass spectrum for each detected feature to provide identification of metabolites. EI ionization operates with electron energies of 70 eV and is a highly reproducible process, which imparts significant internal energy (typically 10–20 eV) to the molecular ion during the ionization process. Covalent bond fission follows with the creation of charged (positive charged ions are most abundant) and neutral fragment species. The charged fragment ions are detected and represented in an EI mass spectrum. The fragmentation process is dependent on the molecular structure, and the resulting mass spectrum provides information for metabolite identification. The mass spectra for glutamic acid and glutamine are shown in Fig. 1. A clear differentiation between the two metabolites is achievable by visual inspection, as the two metabolites have different molecular formulas and chemical structures.

Although widely applied, one limitation of EI mass spectra for metabolite identification, especially when trimethylsilyl (TMS) derivatisation is applied, is the low abundance and in some cases the absence of the molecular ion ($M^{+\cdot}$), and in many cases a M-$15^{+\cdot}$ fragment ion, in the mass spectrum. Defining the mass of the molecular ion is important in de novo metabolite identification to assist in determination of the molecular formula (as discussed in Sect. 4). Chemical ionization approaches (CI or atmospheric pressure chemical ionization; APCI) are useful for detection of (quasi-)molecular ions on GC–MS platforms (Kumari et al. 2011; Wachsmuth et al. 2011).

The EI mass spectrum can be compared to mass spectra acquired from authentic chemical standards accessible in commercially or freely available mass spectral libraries to aid identification. Several libraries are available which

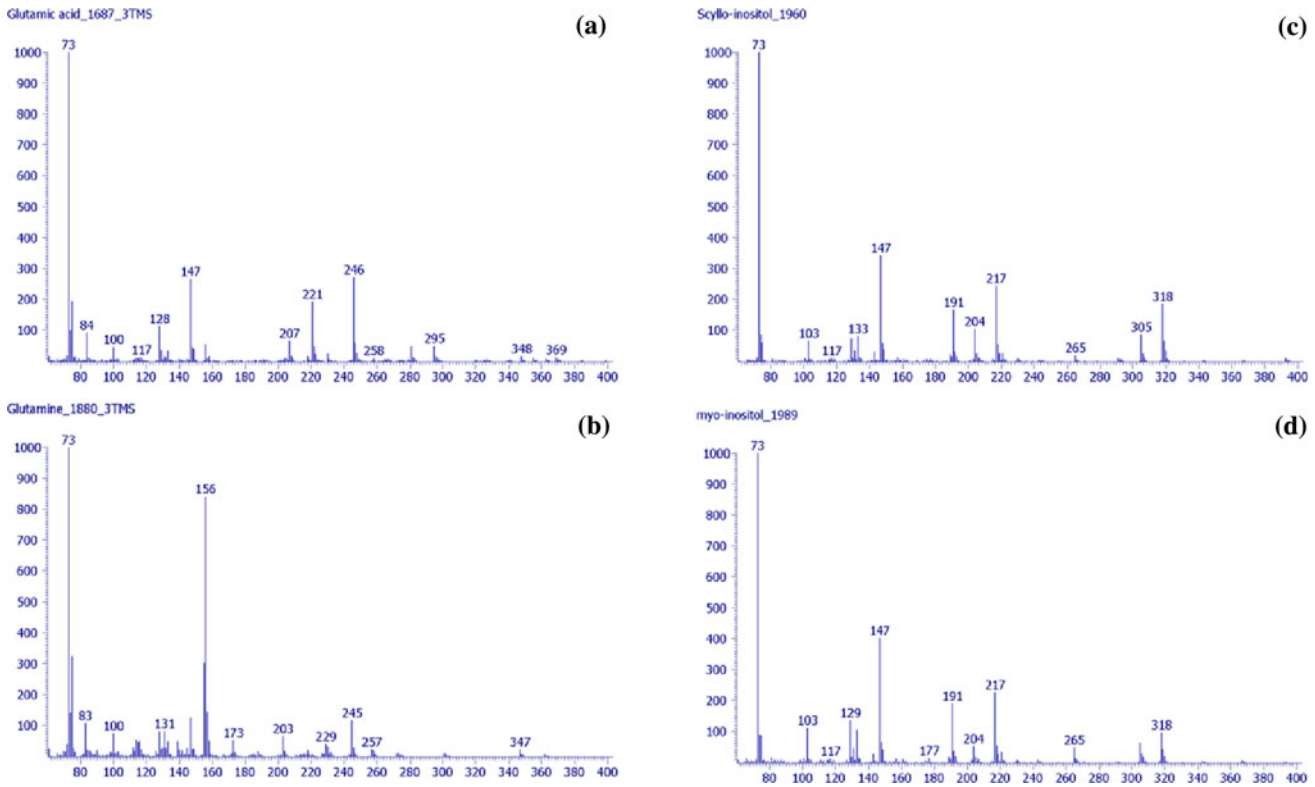

Fig. 1 EI mass spectra for (**a**) glutamic acid, (**b**) glutamine, (**c**) *scyllo*-inositol and (**d**) *myo*-inositol. Clear differences in the EI mass spectra are observed for glutamic acid and glutamine which allow their robust differentiation and identification. No clear difference is observed in the EI mass spectra for *scyllo*-inositol and *myo*-inositol and chromatographic separation is required to provide differentiation

contain either a generalized collection of chemicals [of which a subset are metabolites; for example NIST08 (http://chemdata.nist.gov/mass-spc/ms-search/)] or only metabolites [for example, Golm Metabolome Database (GMD) (Kopka et al. 2005) and FiehnLib (Kind et al. 2009)]. The fragmentation patterns depend on the derivatization process applied; available metabolite-specific libraries mostly contain information on TMS-derivatized metabolites. For other derivatisation processes, research laboratories can construct their own libraries with authentic chemical standards [for example, see Smart et al. (2010) who constructed a library for methyl chloroformate derivatives]. As the EI ionization and subsequent molecular ion fragmentation processes are highly reproducible across different GC–MS platforms they can provide putative (level 2) annotations. In some cases the mass spectra of two different metabolites (most importantly, stereoisomers) are very similar, and identification of a metabolite is not possible at level 2, but classification to a specific metabolite class (level 3) is possible. A common example of this problem is mono- and disaccharides, or other polyhydroxylated metabolites, which occur in multiple different stereoisomeric forms with widely different biological functions (see Fig. 1 for an example of two cyclohexanehexols, *myo*-inositol and *scyllo*-inositol, which have almost identical mass spectra).

To achieve confident identification (level 1), the EI mass spectra information is combined with comparison of chromatographic RTs, or even better Kovats retention indices (Malvoisin et al. 1979; Lisec et al. 2006; Dunn et al. 2011b). Retention indices are a normalized measure of RT that takes into account differences in column length, internal diameter, film thickness, flow rate of carrier gas, and oven temperature ramp, by spiking chemicals from a homologous series into each analyzed sample. This procedure exploits the fact that the RT depends monotonically and reproducibly on the number of carbon atoms among the members of a homologous series for a given chromatographic stationary phase. Retention times of the metabolites of interest are then compared to the RTs of the members of the homologous series and expressed as a retention index relative to the number of carbon atoms in the most similar retention index markers. For example, *n*-alkanes are commonly applied as retention index markers (typically C_{10}–C_{30}) (Lisec et al. 2006; Dunn et al. 2011b). Each of the alkanes is assigned a retention index that is calculated as its carbon number multiplied by 100 (for example, *n*-decane ($C_{10}H_{22}$) has a retention index of 1,000). Then, if C_{10} and C_{12} *n*-alkanes have RTs of 700 and 900 s and an unidentified metabolite has a RT of 800 s, its retention index is calculated as 1,100. Fatty acid methyl

esters have also been applied as retention index markers in metabolomic studies (Kind et al. 2009). The application of retention indices can allow the comparison of data across different GC–MS platforms, though accuracy can be dependent of column dimensions and stationary phase. Retention index data are contained in many mass spectral libraries.

3.2 Novel and developing methods

The identification methods described so far fail for those metabolites not present in mass spectral libraries. As a consequence GC–MS is used mainly for the semi-targeted analysis of known volatiles (e.g. Tikunov et al. 2005) or of ubiquitous primary metabolites (e.g. Fiehn et al. 2000a; Roessner et al. 2000) for which authenticated reference substances can easily be acquired. The large portion of still unidentified metabolites from GC–MS metabolic profiles is in most cases neither analyzed nor reported. These "uncharted" metabolites can typically exceed ~66 % of all mass features detectable by GC–EI–MS profiling of plants (Kopka J., unpublished data), though this estimate is highly dependent on the sample origin; for example, in well studied microbes the number of metabolites identified is higher (van der Werf et al. 2007). Advances in the fast classification and identification of metabolites are essential, as it is obvious that only identified metabolites can be experimentally assessed, manipulated and understood in terms of their physiological role or their involvement in disease mechanisms.

As the first step towards discovery and reporting of so far unidentified metabolites, the concept of mass spectral tags (MSTs) has been introduced (Desbrosses et al. 2005; Kopka 2006). In an analogy to the expressed sequence tags (ESTs) of molecular biology, MSTs represent the physicochemical properties of so far unidentified metabolites. In the case of GC–MS, MSTs comprise typically the full mass spectrum and the chromatographic retention index (e.g. Strehmel et al. 2008) of the chemically derivatized or non-derivatized metabolite. Once these reference data are indexed and archived in public databases, such as the GMD (Kopka et al. 2005) or BinBase/FiehnLib (Kind et al. 2009), targeted searches and the matching of unidentified metabolic features from GC–EI–MS studies of complex samples to such reference MSTs can be performed. The respective processes of matching the so far unidentified metabolite and of the search for MSTs representing known metabolites are essentially equivalent. Indeed GMD has become a central repository of such unidentified MSTs, primarily but not exclusively for plant metabolomic studies, next to the central community function of GMD as an archive of biologically relevant GC–EI–MS reference data of pure and authenticated reference substances (Wagner et al. 2003; Kopka et al. 2005; Schauer et al. 2005; Hummel et al. 2010). In the last 6 years, hundreds of such unidentified MSTs with known relevance to biological samples have accumulated and now await structural elucidation (Kopka 2006; Hummel et al. 2010). The MST concept has also recently been extended towards LC–MS features (Matsuda et al. 2009; Fernie et al. 2011).

Whereas the cataloging of MSTs in metabolomic databases is the necessary descriptive basis for metabolite discovery, efforts towards structural elucidation are increasingly urgent. Still the most successful strategy of metabolite identification is the analysis of pure reference substances on the GC–MS system and the archiving of obtained reference MST data together with the respective structures and, where possible, their retention indices. Such archives can easily be exchanged between laboratories via public and academic databases, such as the GMD (e.g. (Kopka et al. 2005; Strehmel et al. 2008). The commercially available reference metabolites are, however, almost exhausted, and shotgun approaches, which propose to map all commercially available compounds, are too expensive and inefficient because not all biologically relevant compounds are commercially available. Therefore, the chemical synthesis of metabolites, and in principle the biosynthetic production of metabolites by heterologous expression of enzymes with known function will become increasingly important (for an example of biosynthesis of metabolites by heterologous expression, though not for application in metabolite identification see Komatsu et al. 2010).

Alternatively, metabolite identification can employ a direct approach with NMR spectroscopy, i.e. the purification of chemically derivatized or native GC–MS analytes either by preparative GC (Eyres et al. 2008; Ochiai and Sasamoto 2010) or by a combination of preparative LC (Wang et al. 2010) and mapping of the obtained pure fractions to the GC–MS profiling system. Unfortunately, a huge discrepancy can exist between the 10–100 ng amounts which are typically detectable in analytical GC- MS runs (e.g. Birkemeyer et al. 2003) and the approximately 0.1–1 mg required for structural elucidation by NMR. Multiple and replicate injections to collect sufficient amounts of a purified metabolite can be performed. The same forward identification strategy can be applied for LC–MS and CE–MS platforms (Dear et al. 1999). The synthesis or production of xenobiotic metabolites in cell culture systems (Schmidt et al. 2006) and subsequent purification and identification will be an important tool in future human studies, as are informatics approaches to predict detoxification mechanisms of xenobiotics (Kirchmair et al. 2012). Until direct forward identification for MS based metabolomics and other technology platforms become more efficient, approaches that direct the chemical or biochemical synthesis or the purification processes are in high demand.

Several other options for the classification (level 3 identification) of unidentified MSTs exist or are emerging

(e.g. Hummel et al. 2010). Such classifications narrow the structure search space for unidentified metabolites and thus support full structural elucidation and inference of novel biosynthesis pathways. The first gadget in the classification tool box is the determination of the molecular formula of the chemical derivative and/or the native metabolite, as had been suggested more than a decade ago (e.g. Fiehn et al. 2000b). This will be discussed in more detail in Sect. 4. However, systems that couple GC to high mass accuracy (time-of-flight) spectrometers are now available. This will possibly lead to substantial progress for GC–MS profiling. A new generation of GC–MS technology for the identification of metabolites in untargeted experiments would ideally combine (i) sensitive and efficient ionization processes with maintenance of molecular ions (which are essential for metabolite characterization), (ii) MS/MS capability for monitoring fragmentation, and (iii) high mass accuracy for deducing the molecular formula of both molecular and fragment ions. We believe these systems will start to become a more standard and frequently applied tool for enhanced MST classification and structural elucidation.

It is, however, well known that even highly accurate *m/z* measurements can lead to assignment of multiple possible molecular formulas. The number of putative formulas can be reduced by assessing the exact masses of naturally occurring isotopomers and the natural relative isotopomer abundances (also defined as relative isotope abundances). An even more effective approach uses element-specific full in vivo stable isotope labeling, e.g. by ^{13}C or ^{15}N, to increase the signal obtained from the isotopomer masses (Birkemeyer et al. 2005; Kopka 2006; Huege et al. 2007, 2011). Comparison of fully ^{13}C labeled isotopomers and the unlabeled metabolite can be used to determine the number of carbon atoms in the formula of the unidentified MST or EI-fragment. Fully labelled samples can also yield additional exact mass measurements, which allow the reduction and ultimately removal of molecular formula ambiguities by intersection analyses of the respective hit lists. The use of natural and experimentally enriched mass isotopomers for molecular formula inference has been recently demonstrated for the analysis of unidentified metabolites from LC–MS profiles (Giavalisco et al. 2009). Before such approaches towards structure classification can be routinely applied, the available technologies need to be thoroughly evaluated with regard to the current limitations of sensitivity, mass accuracy, and reproducibility as well as their appropriateness for this highly demanding task of metabolite identification.

When experimental advances are limited, it becomes even more important to make the most of existing mass spectral libraries and the linked molecular structure information. Even though not fully understood, EI mass fragmentation patterns reflect the underlying molecular structures and substructures, and careful manual interpretation of the spectra can often elucidate the chemical structure. This traditionally slow and manual process can now be automated, and has been one of the earliest proof-of-concept applications of machine learning technology. One such example is the substructure identification option of the NIST08 mass spectral library (http://chemdata.nist.gov/mass-spc/ms-search/), which is highly useful for the interpretation of EI mass spectra. As an alternative to this 'black box' machine learning approach, GMD uses decision tree technology for the prediction of substructures from GC–MS fragmentation patterns and retention indices (Hummel et al. 2010); this approach has the advantage that the underlying rules of classification are available in a human readable and understandable form. In short, the molecular structures represented within GMD were partitioned into classes that contain or do not contain predefined substructures that frequently occur in metabolites, for example amine-, carboxyl-, carbonyl and hydroxyl-moieties. Decision trees were then trained to predict the presence/absence of these substructures according to abundance thresholds of fragment masses, with retention index thresholds or mass differences within EI mass spectra also contributing in a few cases. These machine learning technologies support molecule identification by predicting the substructures present in unidentified metabolites (with associated confidence/quality scores), accelerating the classification efforts and directing the structure elucidation process (applying MS^n, mass isotopomer analysis and determination of exact mass) and the targeted chemical or biochemical synthesis of the inferred metabolites.

Finally, prediction of RTs or retention indices can also enhance the putative annotation of unidentified metabolites (Mihaleva et al. 2009; Kumari et al. 2011). When combined with accurate mass measurements and elemental formula calculations, this strategy reduced the number of putative molecular formulas (and in some cases returned a single molecular formula (Kumari et al. 2011). The ability to predict metabolite structures and retention indices *in silico* is a significant advancement for the identification of unknown metabolites where data from authentic chemical standards are not available in mass spectral libraries.

4 Metabolite identification in LC–MS, CE–MS and DIMS derived datasets

4.1 The use of accurate *m/z* measurements to define molecular formula and to search electronic resources

The accurate measurement of the *m/z* is frequently the first process applied in the chemical identification of metabolic

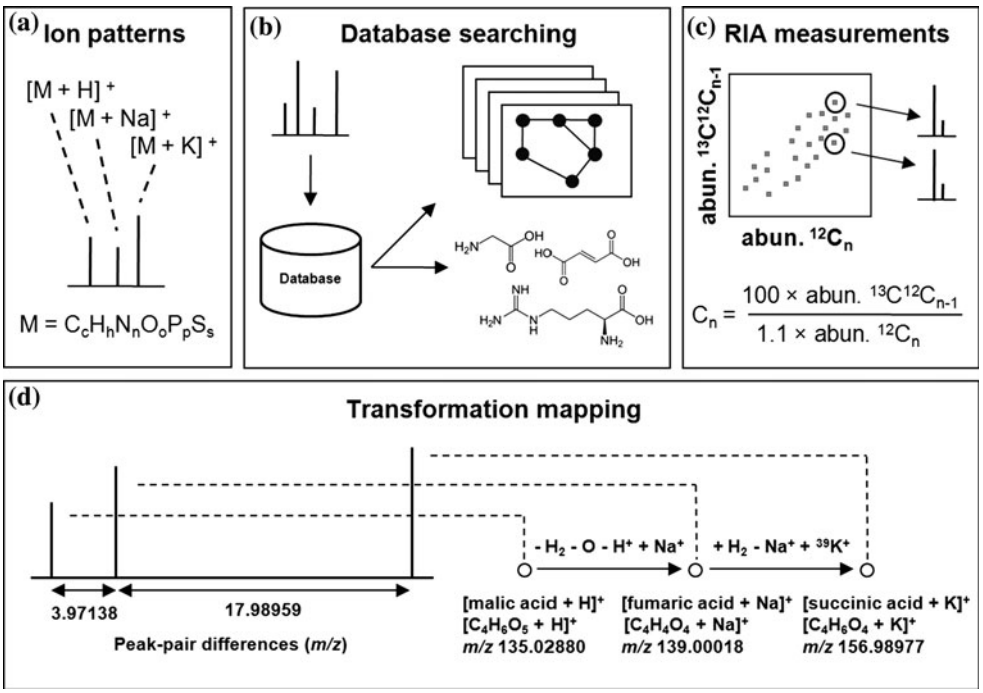

Fig. 2 (**a**, **b**) The application of chemical, biological and MS data to metabolite annotation of DIMS, LC–MS, UPLC–MS and CE–MS data. A mass spectrum typically comprises of hundreds or thousands of signals (or 'features') arising from many different metabolites as well as the naturally occurring isotopes (e.g., $^{13}C^{12}C_{n-1}$), adducts and fragments of these metabolites. Putative metabolite annotation of these signals can be in principle achieved using accurate measurements of *m/z*; typically by assigning one or more molecular formula (i.e. $M = C_cH_hN_nO_oP_pS_s$) to each accurate *m/z* measurement or searching against a compound database on a peak-by-peak basis. (**c**) RIA measurements can be used to determine the numbers of certain atoms (e.g. C) present in a metabolite, which ultimately improves the accuracy of assignment. (**d**) Biological samples comprise of thousands of metabolites that are related through specific chemical transformations or networks. Prior biological knowledge of these transformations, in the form of substrate-product pairs, can also significantly increase the accuracy of metabolite identification

features detected in data sets acquired on LC–MS, CE–MS and DIMS platforms. These platforms typically utilize ESI sources (Fenn et al. 1989), though other ion sources (or combinations of ionization mechanisms) are sometimes used, including APCI (An et al. 2010). Measurements of *m/z* can be used to match a metabolic feature to a single or small number of molecular formula. The accuracy of this measurement defines the number of molecular formula matches; the greater the accuracy the lower the number of molecular formula matches. The majority of mass spectrometers applied to this task operate at high mass resolutions (5,000 to greater than 200,000) and mass accuracy (<5 ppm) and include TOF and Fourier Transform-based instruments but not quadrupole and ion trap instruments. The molecular formula or formulas are then matched to metabolites via searching of on-line databases. A single molecular formula can correspond to multiple known metabolites; therefore the application of accurate measurements of *m/z* is an appropriate first step, but only provides putative (level 2 or 3) annotation requiring further verification. Other chemical and biological knowledge can also be applied, in parallel to or in combination with accurate measurements of *m/z*, to limit the number of putative metabolite annotations for a single metabolic feature. These will be discussed in the next section. The workflow of applying accurate *m/z* data and biological knowledge is summarized in Fig. 2.

The process of matching *m/z* to molecular formulas starts with a large search space composed of all potential molecular formulas. The reduction in search space size (or in the number of potential molecular formulas) is achieved by the matching of experimentally derived *m/z* information to the equivalent *m/z* of specific molecular formulas. The efficiency of this reduction process is dependent on the resolution and accuracy of the mass spectrometer. The majority of mass spectrometers operate with mass resolutions of 5,000–50,000, which allows the resolution of metabolites with the same nominal mass but different monoisotopic masses [for example, glutamine (monoisotopic mass = 146.0691) and lysine (monoisotopic mass = 146.1055)]. As mass resolution increases, the ability to resolve ions of the same nominal mass but different monoisotopic mass is increased (see Fig. 3 for an example). It should be remembered that even with high

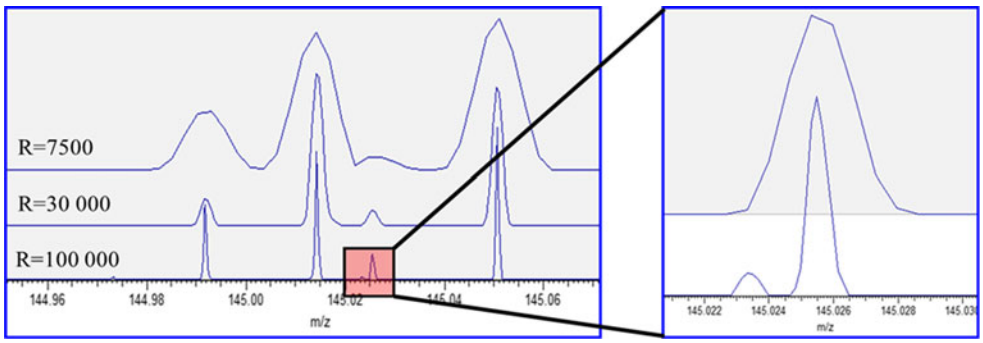

Fig. 3 The advantage of high mass resolution for the differentiation of metabolic features with similar *m/z*. The data were acquired on a ThermoFisher LTQ-Orbitrap XL hybrid mass spectrometer operating at mass resolutions (FWHM) of 7,500, 30,000 and 100,000. The sample analysed is human serum, typical of metabolomics experiments

mass resolution, high mass accuracy is not necessarily achieved and appropriate mass calibration is required to provide high mass accuracy in these measurements (e.g., Scheltema et al. 2008). As mass accuracy increases, the mass error range decreases and the number of proposed molecular formulas decreases. This relationship is dependent on *m/z*, and as *m/z* increases so will the number of possible molecular formulas matching a given mass for a defined mass error (Kind and Fiehn 2006). In this process, unambiguous determination of a single molecular formula is not always achievable even with high mass resolution and the achievement of sub-ppm mass accuracies (Kind and Fiehn 2006).

In LC–MS, CE–MS and DIMS applications, thousands of metabolic features are detected, defined as *m/z*–RT pairs (except for DIMS where no RT data are available). Significant complexity is entwined in these data. A single metabolite is typically (but not always) detected as multiple metabolic features, each having the same RT but a different *m/z*. The different *m/z* values relate to different derivative ions of the same metabolite. These can include protonated and deprotonated ions, adducts, fragments, isotopomers, dimers, multiply charged ions and Fourier transform artifact peaks (Brown et al. 2009). The derivative ions depend on the chemical properties of the metabolite as well as on the sample matrix, solvents, metabolite concentrations, and mass spectrometry platform and parameters (Tong et al. 1999; Zhu and Cole 2000; Schug and McNair 2002, 2003; Brown et al. 2009). High salt contents can lead to complex gas-phase non-covalent interactions during ESI ionization; metabolite cluster ions containing multiple salt ions, including Na^+, K^+ and Cl^-, have been observed in blood and urine (Brown et al. 2009). Sample preparation can be used to reduce this complexity, for example applying a desalting process, though desalting of samples can lead to unwanted loss of metabolites and therefore is not usually applied in untargeted analysis. Figure 4 shows an example of the complex mass spectrum detected for a single, compound on a sensitive high-resolution mass spectrometer.

The wide variety of different ion types detected can be applied advantageously if appropriate methods are used, but can provide great difficulties and significant errors if ignored. The complexity of ESI data can lead to a large number of false positive identifications, especially when derivative ions are falsely identified. For example, in Fig. 4, a range of metabolic features, relating to different ion types of tyrosine are shown. If the $[M + Na + Na]^+$ adduct of tyrosine is labeled as a protonated ion then accurate metabolite identification will fail as an incorrect molecular formula will be calculated.

The automated matching of metabolic features deriving from the same metabolite (including the automated determination of the ion type) has only recently been applied in metabolite identification. Metabolic features derived from the same metabolite are identified, annotated, and grouped together using accurate *m/z*, *m/z* differences, RT similarity, pairwise correlation between measured responses, known adduct lists and chromatographic peak shape similarity. Metabolite identification for one of the metabolic features can then be linked to all other derived metabolic features for that metabolite. The freely available software platforms developed for this purpose include PUTMEDID-LCMS (Brown et al. 2011), CAMERA (Kuhl et al. 2011), PeakML/mzMatch (Scheltema et al. 2011) and IDEOM (Creek et al. 2012). These software platforms can be applied, in most cases, to data acquired on different analytical systems and pre-processed by different software packages. The applicability to diverse datasets can be enhanced by developing sample-specific reference files containing molecular formula and metabolites for specific sample types. These reference files can be organism-specific to allow an appropriate reduction of the chemical search space before identification processes are performed.

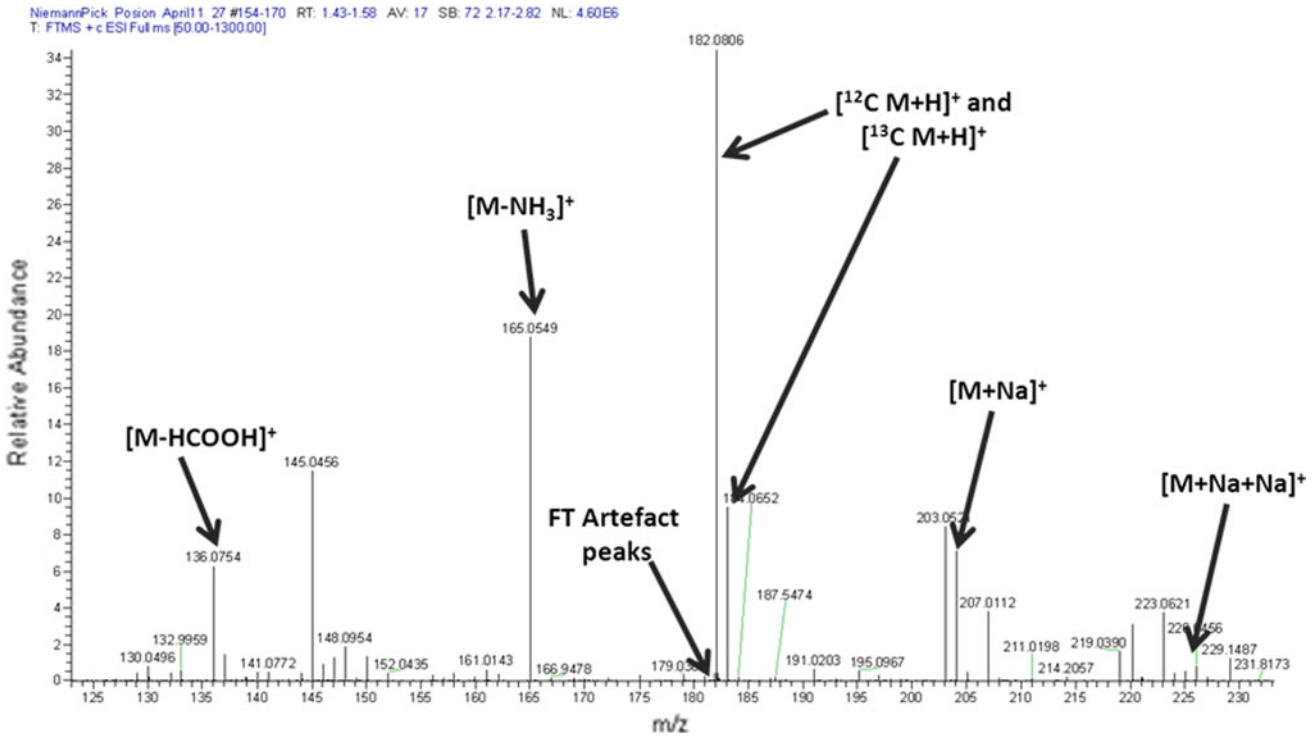

Fig. 4 Full-scan mass spectrum related to the detection of tyrosine on a UPLC–MS system (Waters Acquity UPLC system coupled to a ThermoFisher LTQ-Orbitrap XL hybrid mass spectrometer). The complex array of ion types detected includes loss of formate and ammonia (fragmentation in the ion source), ^{13}C isotopologues and complex salts. FT artifact peaks observed for the $[M + H]^+$ metabolic feature of tyrosine are also present but at such a low intensity so as not to be visible in this figure

The limitations of matching and integration of metabolic features should be investigated before routine application because errors can occur. The *m/z* difference between different metabolic features deriving from the same metabolite (peak-pair *m/z* differences) is routinely used and its applicability has recently been investigated. A method was reported to calculate *m/z* differences across a mass spectrum using commonly occurring peak-pair differences (Weber and Viant 2010). The resulting *m/z* difference error surface, representing the error associated with *m/z* differences between peak pairs (e.g. $[M + Na]^+$–$[M + H]^+$) of the same metabolite as a function of both the mass difference and the average *m/z* of the peak-pair, revealed large relative errors (>100 ppm) for closely mass spaced peaks (Weber and Viant 2010). Hence large error tolerances may be required when analyzing peak differences to avoid false negative assignments, though these are expected to be instrument-dependent.

Following the identification of single or multiple molecular formula, the matching of these molecular formulas to known metabolites is performed, typically by searching an array of online or laboratory-specific resources [including HMDB (Wishart et al. 2009), KEGG (Ogata et al. 1999), LipidMaps (Sud et al. 2007), PubChem (http://pubchem.ncbi.nlm.nih.gov/), and ChemSpider (http://www.chemspider.com/)]. It is important to define a valid molecular formula first and then match a molecular formula to a metabolite. In cases where a match to a metabolite is not possible, the molecular formula can then be applied in subsequent identification processes.

4.2 Application of chemical, biological and other experimentally derived MS data

As discussed above, the majority of *m/z* measurements in a single complex biological mass spectrum cannot be assigned to a single molecular formula based on accurate measurements of *m/z* only (Kind and Fiehn 2006; Kind and Fiehn 2007; Matsuda et al. 2009). A range of bioinformatics approaches have been developed in the past few years to exploit relationships between signals in high-mass resolution mass spectra. These include peak-pair *m/z* differences (for example, peak *m/z* patterns and prior biological knowledge) and peak-pair intensity ratios (for example, isotope abundance ratios and peak area correlations) as tools to reduce the number of putative molecular formulas or metabolite assignments for a single metabolic feature and thus aid in its identification. Other tools developed include ionization behavior rules, applying chemical knowledge of metabolites, which can be used to determine the probability of the formation of specific

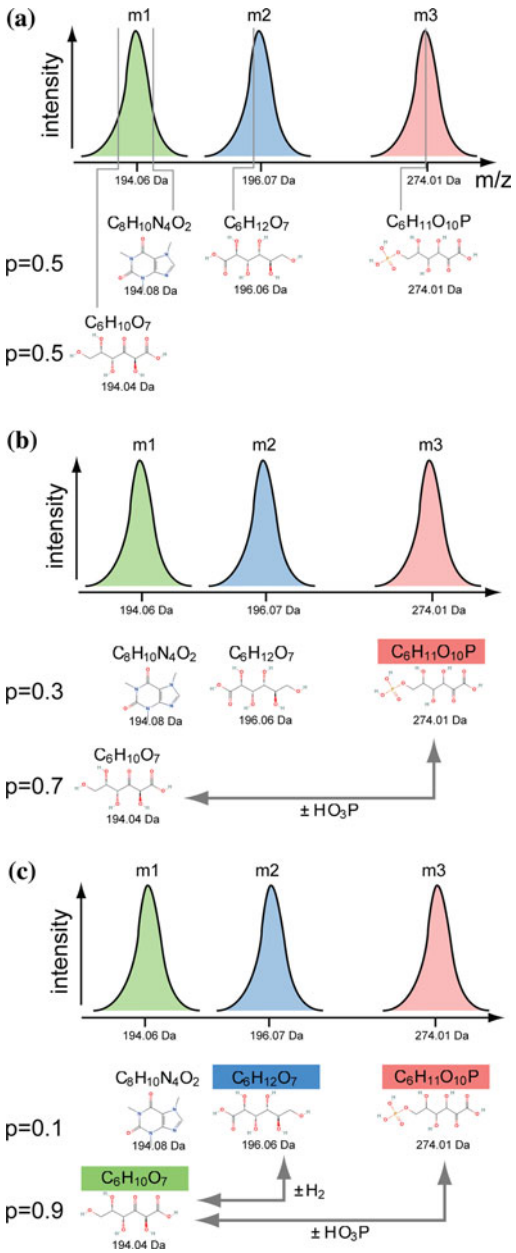

Fig. 5 The principle of using biological knowledge in the form of putative enzymatic transformations to improve the confidence in metabolite identifications. (**a**) Three metabolic feature have been observed in this example; for two of them, m_2 and m_3, the identification is unambiguous, based on the mass alone. For the third one, m_1, two possible formulas match the peak, both being initially equally probable. (**b**) Observing that one of the possible formulas, $C_6H_{10}O_7$, is linked to an unambiguously identified metabolite via a putative dephosphorylation, increases the posterior belief in this identification, as indicated by the increased *p*-value (the exact *p*-value will depend on the relative weighting of the two sources of evidence, *m/z* and *m/z* differences). (**c**) Once a second putative enzymatic relationship, a dehydrogenation, is detected, the preference for the identification as $C_6H_{10}O_7$ becomes a near certainty. The Bayesian algorithm described in Rogers et al. (2009) performs this analysis simultaneously for the entire set of observed metabolites, and identifies the most plausible set of identifications, as well as possible alternative interpretations of the dataset

derivative ions for a metabolite and to eliminate metabolites with chemically infeasible ion types (Draper et al. 2009).

Prior biological knowledge can be applied to constrain the metabolite search space and aid metabolite identification, e.g. using reference lists of expected metabolites for specific sample types or a specific study organism. Although strict search parameters have been shown to improve the accuracy of molecular formula annotation (e.g. narrowly constraining the mass error tolerance, elements allowed, numbers of each element allowed, ion types etc.), additional bioinformatics approaches are necessary to further increase identification accuracy (Kind and Fiehn 2006, 2007). Biological samples are not composed of random metabolite mixtures, but instead comprise of thousands of biochemically related compounds resulting from the loss and/or gain of atoms between substrate–product pairs (Breitling et al. 2006a, b). Integrating prior biological knowledge in the form of such enzymatic transformations into metabolite identification has proven successful at reducing the number of false structural and non-structural assignments (Gipson et al. 2008; Rogers et al. 2009; Weber and Viant 2010). For example, the mass error surface, based on detected *m/z* differences, together with the inclusion of prior biological knowledge from the KEGG database (Kanehisa et al. 2010) has been shown to decrease the false positive rate of metabolite identification by more than fourfold (Weber and Viant 2010). A similar method that incorporated LC RT measurements has likewise demonstrated an increased confidence in metabolite identification (Gipson et al. 2008). These methods are dependent on prior biological knowledge, for example the metabolic network of the study organism. This is not always available, especially when studying obscure biological systems or diverse areas of biology distant from central metabolism. An example of how to apply biological knowledge in the form of putative enzymatic transformations is shown in Fig. 5.

RIA measurements can be applied to reduce the chemical search space even further, and are applied routinely with data acquired on instruments where accurate isotope abundance measurements are possible, providing an estimate of the numbers of specific atoms present in a particular parent peak (Xu et al. 2010; Weber et al. 2011). This is highlighted in Fig. 2. Several theoretical and experimental studies have shown the benefit of using RIA measurements to remove incorrect empirical formula assignments (Kind and Fiehn 2006; Kaufmann 2010; Miura et al. 2010). Furthermore, several studies have characterized the accuracy and precision of RIA measurements on different MS platforms (Stoll et al. 2006; Koch et al. 2007; Erve et al. 2009; Xu et al. 2010). The higher the accuracy and precision of RIA measurements the more incorrect empirical

formulas assignments can be discarded; however, even with relatively inaccurate and imprecise RIA measurements the annotation of metabolic features can be improved significantly (Weber et al. 2011).

Stable ^{13}C isotope labeling (as described above for GC–MS) is another technique for which the principle of peak-pair *m/z* differences has been exploited successfully to distinguish background ions from ions of true biological origin; the proof of principle study applying GC–MS identified over 1,000 formulas of biological origin and reduced the number of false positive molecular formula assignments (Giavalisco et al. 2008). Subsequently, this method has been improved using UPLC–MS to increase the accuracy of identification, i.e. to achieve structural identification and accurate relative quantification (Giavalisco et al. 2009). Furthermore, an extension of this approach using dual stable isotope labeling (i.e. $^{13}C^{12}C_{n-1}$ and $^{15}N^{14}N_{m-1}$) of metabolites, instead of single labeling, has been shown to be a valuable tool for discovering novel chemical structures (Feldberg et al. 2009).

Signal intensities can also be used to improve metabolite identification using, for example, the linear correlations between signals of specific peak-pairs measured across multiple mass spectra. This is applied to the annotation of molecular and derivative ions as described in the previous section (Iijima et al. 2008; Draper et al. 2009; Brown et al. 2011; Weber et al. 2011). Specifically, high correlation coefficients (e.g. arbitrary threshold of $R > 0.9$) for these intensity relationships have been used previously to increase the confidence of assigning frequently detected *m/z* differences (Iijima et al. 2008; Brown et al. 2009, 2011; Fuhrer et al. 2011). Such strongly correlated relationships occur in particular for isotopologues (e.g. $^{12}C_n$ and $^{13}C^{12}C_{n-1}$), as their intensity relates directly to their natural abundance.

The majority of bioinformatics approaches and methods described in the previous section and in this section are focused primarily on the two principal variables measured in a typical MS experiment, *m/z* and signal intensity. In many metabolomics studies, chromatographic or electrophoretic separation of the complex biological mixtures prior to MS analysis is routine (De Vos et al. 2007; Lu et al. 2008; Kenny et al. 2010). The RT or migration time is predictive of metabolite structure, primarily hydrophobicity, hydrophilicity and/or charge for LC–MS, and charge and cross-sectional diameter for CE–MS. For example, in reversed phase LC, hydrophilic metabolites elute at earlier RTs compared to hydrophobic metabolites. In GC–MS, RTs or indices are routinely applied across different platforms to aid identification by comparison of MS and RI data to mass spectral libraries (Kopka et al. 2005), but in LC–MS retention behaviors are far less reproducible and their use for identification much more restricted; changes in LC column phase (for example, reversed phase compared to HILIC), the manufacturer (or sometimes the batch) of columns of the same stationary phase, solvents and gradient elution conditions influence the RT, in many cases significantly. Therefore, although LC–MS focused mass spectral libraries or databases are available [for example, METLIN (Sana et al. 2008) and MassBank (Horai et al. 2010)], comparison of RT data is only applicable for data acquired with the same analytical method on the same equipment and columns. Mass spectral libraries applying RT data are not as readily transferable as is observed for GC–MS applications. Therefore, as most researchers apply different methods and equipment, the ability to apply mass spectral libraries across the research field in many different research groups is low. To improve the usefulness of specific LC–MS libraries, common, standardized analytical methods and instruments would need to be applied. However, as for GC–MS, the ability to predict RTs or migration times is an important recent research area to further reduce the number of potential metabolite identifications. Retention time or migration time prediction has been shown to be achievable for HILIC–MS (Creek et al. 2011) and CE–MS (Sugimoto et al. 2005), respectively.

4.3 Application of experimentally derived MS/MS and MS^n data

The previous two sections describe processes applied to reduce what can be an exceptionally large search space of molecular formulas down to a single or small set of metabolites (or molecular formulas). These processes operate well in reducing the search space size but do not necessarily lead to unambiguous (level 1) identifications. In many cases, multiple possible metabolites are reported for a single metabolic feature, in particular for stereoisomers. This is a significant issue for carbohydrate research and the metabolomics study of complex lipid samples, where each lipid can be composed of different combinations of fatty acids which correspond to the same molecular formula. For example, a diacylglyceride (DG) with a molecular formula of $C_{43}H_{78}O_5$ will contain two fatty acid side chains with a total of 40 carbon atoms and three unsaturated bonds, but these can be distributed in many different structural arrangements, such as DG(20:1/20:2) and DG(18:0/22:3), and many of these are actually found in biological samples. Also, the structural position of each fatty acid or unsaturated bond, as well as its stereochemistry (*cis*/*trans*) can be biologically important, and defining these small structural differences is a necessity. To provide further data for structural elucidation or to aid in de novo structure elucidation (where no matches to known metabolites were observed applying the techniques described above), fragmentation of molecular ions is applied, with detected mass

spectra being dependent on the structure of metabolites. This process has different levels of efficiency depending on the metabolite class. Classes of lipids are composed of similar building blocks with only minor differences in, e.g., fatty acid chain length and double bond number and position. MS/MS fragmentation can lead to characteristic fragment ions for each unit to aid identification. For example, glycerophosphocholines can lead to fragment ions characteristic of the phosphocholine head group and the two fatty acid moieties and even the position of a double bond (Castro-Perez et al. 2011). For this reason systematic identification of lipids can be simpler than for other metabolites. Gas phase fragmentation has been discussed previously (for example, see de Hoffmann and Stroobant 2007 and Kind and Fiehn 2011) and will only be briefly described here.

Commonly, three different gas-phase ion activation strategies are applied with LC–MS or CE–MS metabolomic platforms. These are (i) collision induced dissociation (CID) in an ion trap (IT) platform, (ii) CID in a quadrupole-time of flight (Q-TOF) or triple quadrupole (QQQ) platform, and (iii) higher energy collision dissociation (HCD) in Orbitrap instruments. Ion activation applying CID in a Q-TOF or QQQ instrument operates by acceleration of ions through a collision cell separating two other mass analysers and containing a higher pressure of a gas (typically nitrogen or argon). Ion–gas molecule collisions occur in the collision cell imparting internal energy to the molecular ion. CID in an ion trap instrument operates by acceleration of orbiting ions resulting in ion-gas molecule collisions and increases in molecular energy. HCD operates in a similar manner to CID in a QQQ instrument by acceleration of ions into an octopole containing an elevated gas pressure.

All of these processes lead to an activation of ions by increasing their internal energy and the subsequent loss of internal energy through the fission of covalent (and when adduct or cluster ions are studied, non-covalent) bonds. Weaker bonds are more likely to break; thus, the type and strength of different covalent bonds in a metabolite will lead to a specific structure-defined fragmentation pattern. The resulting fragments can either retain the ion charge (and therefore be detected by mass spectrometers) or be neutral species (not directly detectable). The different ion activation mechanisms can provide different ion fragmentation mechanisms and different fragmentation mass spectra. An example is shown in Fig. 6 where fragmentation mass spectra have been acquired for decanoic acid applying CID in a linear ion trap and HCD in an Orbitrap instrument.

Tandem mass spectrometry is applied in these processes. For most experiments the process is two-stage and provides a fragmentation mass spectrum for a chosen ion (defined as the precursor or parent ion). This is known as MS/MS or MS^2. In certain cases, when ion trap systems are used, additional levels of fragmentation can be applied, defined as multi-stage mass spectrometry or MS^n, where n is the number of successive fragmentation experiments. Here a precursor ion (or, in classical terminology, a parent ion) can be fragmented in an MS^2 (or MS/MS) experiment, followed by fragmentation of one or more fragments ions (in classical terminology defined as daughter ions) in an MS^3 experiment, followed by fragmentation of one or more granddaughter ions in an MS^4 experiment and so on. Figure 7 shows an example of MS^3.

The processes of MS/MS and MS^n can be important when attempting to discriminate metabolites of the same molecular formula and similar chemical structure where, for example, the type or position of a fatty acid is the only difference between two different metabolites. This process has been described by a number of researchers and has shown great potential in plant and mammalian applications. Different methods to acquire appropriate MS/MS or MS^n data have been developed and applied. MS^E was introduced on Waters instruments in 2006 to acquire extensive MS/MS data during accurate mass full-scan data acquisition (Plumb et al. 2006). Here, alternate full-scan, low collision energy and high collision energy MS/MS scans are acquired to maximize MS/MS data acquisition. Data independent MS/MS data are acquired, all precursor ions present are selected for the MS/MS experiment in comparison to data-dependent experiments where only specific precursor ions are selected for MS/MS experiments. The lack of precursor ion selection can complicate spectral interpretation in complex metabolomics samples. MS^n has been applied to provide greater specificity in the identification process (Sheldon et al. 2009). For example, for MS^3, a precursor ion is fragmented to produce a number of product ions (graphically described in a mass spectrum) followed by fragmentation of each of the product ions which are also represented in a mass spectrum for each product ion fragmentation experiment. The set of fragmentation mass spectra can be represented in a mass spectral tree. MS^n has been applied in the structural elucidation and identifications of polyphenols in plants (van der Hooft et al. 2010). A two-stage CID system integrated with ion mobility mass spectrometry has been applied to define the location of fatty acyl components and the position of double bonds in these components (Castro-Perez et al. 2011) in lipids from mammalian plasma. The application of HCD has been shown to be appropriate for the characterization of mammalian lipidomes (Bird et al. 2011).

MS/MS or MS^n data can be acquired in the same analytical run as accurate *m/z* full scan profiling (on-line MS/MS), by signal-dependent precursor ion selection (MS/MS)

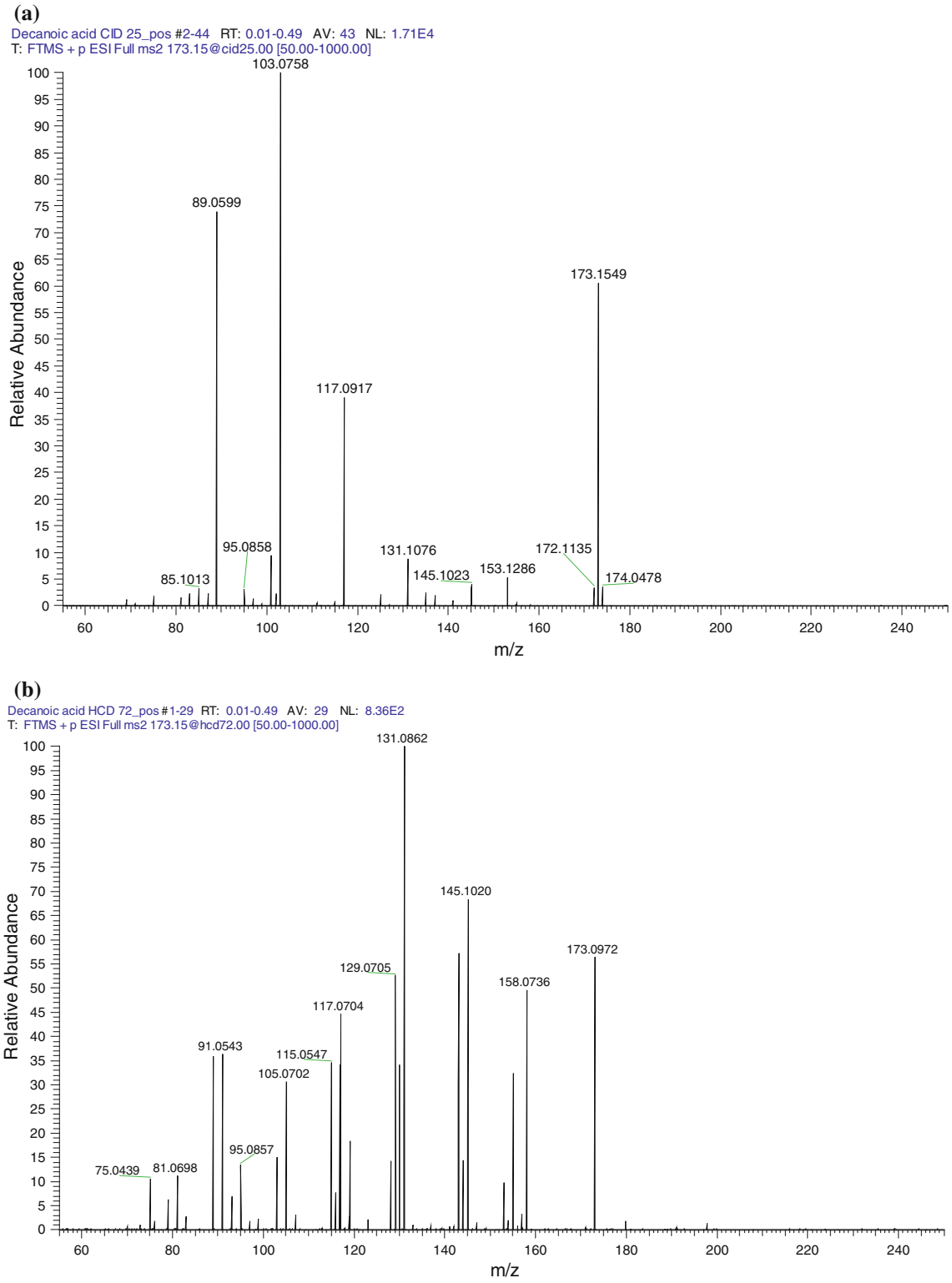

Fig. 6 MS/MS mass spectral differences can be observed depending on the ion activation mechanism applied. MS/MS mass spectra acquired for decanoic acid in (**a**) an LTQ-Orbitrap mass spectrometer with HCD and (**b**) a linear ion trap with CID. The mass spectra for CID and HCD ion activation differ significantly

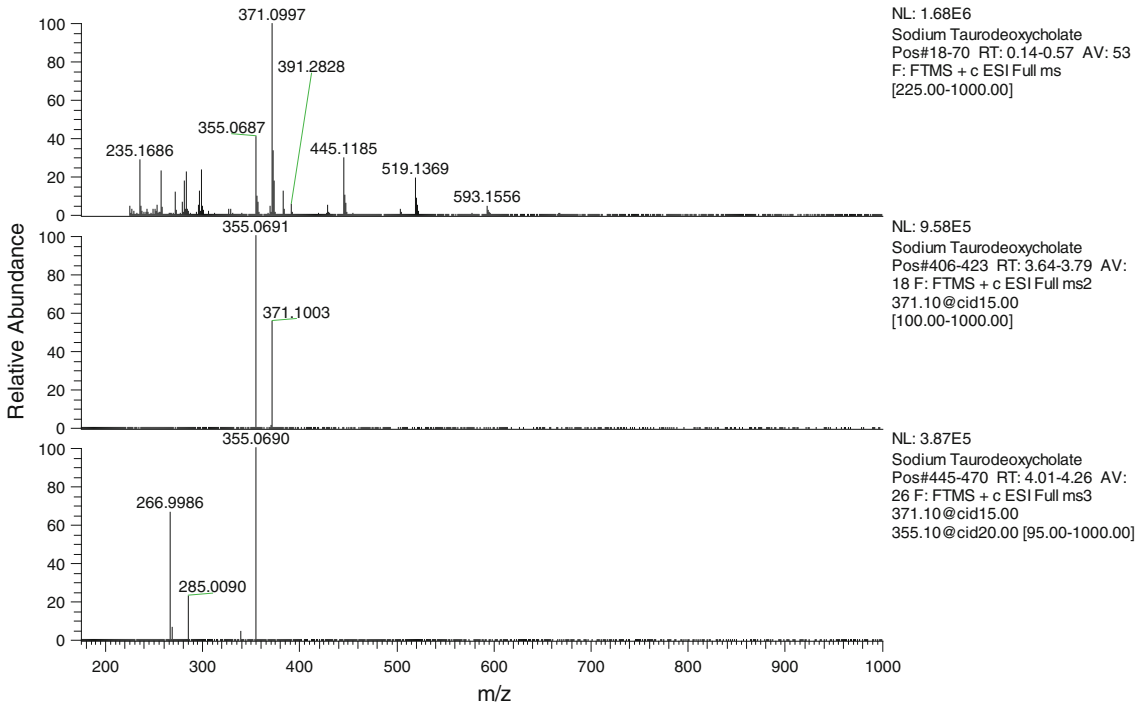

Fig. 7 An example of MS^3 applied to the analysis of taurodeoxycholic acid applying CID ion activation in a linear ion trap and detection in an LTQ-Orbitrap XL mass spectrometer. The *top panel* is a full-scan mass spectrum, the middle panel is a MS^2 mass spectrum of the molecular ion (*m/z* 371.1) and the bottom panel is a MS^3 mass spectrum from the CID fragmentation of a single product ion (at *m/z* 355) produced in the MS^2 experiment

or non-selective fragmentation of all ions (MS^E) or a two-stage CID system integrated with ion mobility mass spectrometry. These approaches apply the same mass spectrometer parameters to all metabolites where MS/MS data are acquired, which is usually not optimal. The optimal MS/MS parameters for one metabolite will not be consistent for all metabolites. The application of different MS/MS parameters across an analytical batch containing up to 100 samples, where one set of unique MS/MS parameters is applied for each sample injection appears to be more optimal (Warwick Dunn, unpublished data). When off-line optimization of collision or activation energies is not possible, as would be performed for single component mixtures where an authentic chemical standard is available, the application of alternating collision or activation energies can be appropriate to maximize the probability of acquiring an information rich tandem mass spectrum. This has been described for HCD experiments performed on mammalian lipidomes (Bird et al. 2011). The acquisition of MS^n data (where $n > 2$) in-line during metabolic profiling can be challenging, particularly in relation to the time needed to acquire MS^n data (typically seconds). An alternative is to perform fraction collection followed by direct infusion of fractions into a nano-electrospray system which can provide minutes of MS^n data acquisition time with sample volumes of less than 10 µl (van der Hooft et al. 2010). Another solution is the replicate analysis of a single sample and the construction of appropriate inclusion and exclusion lists to provide MS/MS data for a larger fraction of the detected metabolome than is possible applying on-line MS/MS with full scan profiling. This has been shown to be successful in proteomic applications (Hoopmann et al. 2009) and is being trialed for metabolomic applications (Neumann et al. 2012).

MS/MS and MS^n data can be compared to data available in mass spectral libraries. Currently, this is less frequently applied in LC–MS than in GC–MS applications, for a number of reasons, as will be described in Sect. 6. MS^n experiments can also be applied to deduce valid molecular formulas. The accurate *m/z* measurement of molecular and fragment species (neutral and charged) can provide improved reductions in the chemical search space and increased confidence in molecular formula determination of molecular ions (Konishi et al. 2007). Several appropriate tools have recently been developed (for example, Rojas-Chertó et al. 2011).

4.4 The use of *in silico* fragmentation tools to aid metabolite identification

The lack of comprehensive metabolite data in GC–MS and LC–MS mass spectral libraries limits the ability to identify

metabolites through mass spectral library searches. However, in many cases fragmentation mass spectra are (or can be) experimentally acquired for these unidentified metabolites, as described in the previous section. These data are available to aid the identification process, and in GC–MS applications this has been applied for sub-structure searching (Hummel et al. 2010). In LC–MS, *in silico* fragmentation software tools are available to enable the matching of *in silico* derived mass spectra (instead of mass spectra derived from authentic chemical standards) to the experimentally derived mass spectra. Typically, a reduction in the search space is performed applying accurate *m/z* and other measurements followed by *in silico* fragmentation of the proposed metabolites. This strategy has been applied successfully in protein studies to construct databases containing data on trypsin-associated cleavage and MS/MS mass spectra of peptides [for example, MASCOT (http://www.matrixscience.com/home) and SEQUEST (http://fields.scripps.edu/sequest/)]. However, the prediction of fragmentation mechanisms for proteins and peptides is significantly simpler than for metabolites, due to the repetitive structure of the linear backbone.

In silico fragmentation tools attempt to construct a fragmentation pattern and associated mass spectrum with regard to a known molecular structure. The *in silico* derived mass spectrum can be compared against an experimentally derived mass spectrum to ascertain whether the metabolite identification is correct. This comparison is based on *m/z* only and not intensity differences. This process can be straightforward for simple compounds, but fragmentation reactions in a tandem MS (or multi-stage MS^n) experiment can exploit the full known (and unknown) complexity of gas phase chemistry. The early approaches towards computer-aided structure elucidation (CASE) were published over 20 years ago. The ASES/MS system was designed for low-resolution GC/MS spectra and combined a library search, association between peaks and substructures, molecular structure generation from building blocks and spectra prediction (Zhudamo et al. 1988). The result was a ranked list of candidates, and thanks to the structure generation step the list was not limited to already known compounds. Each of the modules of the ASES/MS system was rather limited, but the general architecture is still valid. MASSPEC was a system to aid the human expert in interpreting a spectrum with a putative structure in mind (Siegel and Gill 1990), where expert knowledge in the form of "superatoms" of unfragmentable substructures was required. Later systems such as EPIC (elucidation of product ion connectivity (Hill and Mortishire-Smith 2005) did not require such explicit knowledge. However, a common problem is that often, many substructures are able to explain a fragment mass. The Fragment Identification program (FiD) attempts to select the correct structure by removing those bonds with high bond dissociation energies (Heinonen et al. 2008).

Several commercial tools exist for the interpretation of tandem mass spectra or *in silico* fragmentation of metabolites. Both the ACD Fragmenter (http://www.acdlabs.com/products/adh/ms/ms_frag/) and MassFrontier (http://www.highchem.com/massfrontier/mass-frontier) can create a (putative) interpretation of a mass spectrum, and use much more sophisticated fragmentation rules than ASES/MS. For MassFrontier fragmentation mechanisms are based on curated literature data.

Two approaches, developed in academia and freely available, have been published to search general-purpose compound libraries, based on the results of *in silico* fragmentation tools to provide candidate lists of putative metabolite annotations. Hill and colleagues used scripting on top of MassFrontier to produce a ranked list of PubChem compounds applied for metabolite identification (Hill et al. 2008). MetFrag is an open source system consisting of a Java library, command line tools and a web front-end to search KEGG, PubChem or ChemSpider and provide putative candidate lists of metabolites (Wolf et al. 2010). If such systems are to be used on compound libraries as large as PubChem (as of 20th March 2012, PubChem contained more than 32 million entries), the runtime per candidate becomes important. While Hill et al. report 2.5 s per compound, MetFrag requires only 0.2 s per candidate. Together with an even faster (albeit less accurate) candidate pre-selection, (Hildebrandt et al. 2011), that creates peak-to-structure associations in a training step and performs a preliminary ranking directly in a relational database, large numbers of candidates can be evaluated. In addition to databases of known metabolites, the identification can also be performed for purely hypothetical structures obtained through structure generation programs (Schymanski et al. 2011, 2012). These tools require further assessment and validation on different instrument and sample types to define their capabilities and aid in further development. Further advances in the development of *in silico* mass spectral libraries will undoubtedly fill gaps in mass spectral libraries constructed with authentic chemical standards.

5 Metabolomic databases and mass spectral libraries

The available metabolome-focused databases are increasing in both number and size and aid the matching of accurate *m/z* measurements and molecular formulas to metabolite identities. However, it should always be remembered that our current level of knowledge of metabolites present in sample-specific metabolomes is not complete and many features can relate to both previously

unknown endogenous metabolites, and exogenous metabolites from many sources and deriving from diet (Lloyd et al. 2011), lifestyle (Pechlivanis et al. 2010), pharmaceuticals (Loo et al. 2012) and gut microflora (Wikoff et al. 2009). Without the presence of all these metabolites in organism-, tissue- or cell-specific databases, or mass spectral libraries, the inclusion of false positive and false negative identifications is inevitable. Only a small percentage of all known metabolites are available commercially to be incorporated in mass spectral libraries (Brown et al. 2009; Wishart 2011). Therefore mass spectral libraries containing data for all metabolites are unlikely to be ever constructed.

For these reasons care should always be taken when building biological conclusions on putatively annotated metabolites. The putative annotation of *multiple* metabolites shown to be of biological importance and related by metabolite class (e.g., sugars or glycerophosphopholipids) or metabolite pathway (e.g., glycolysis) provides improved confidence that these metabolite classes or pathways are indeed involved in the biological process under study.

For some aspects of metabolite identification, applying accurate *m/z* measurements for example, databases are required which contain purely chemical information. In addition to the general chemical databases, like PubChem (http://pubchem.ncbi.nlm.nih.gov/) or ChemSpider (http://chemspider.com/), others limit their scope to known metabolites either being species-specific or species non-specific [including KEGG (Ogata et al. 1999), MetaCyc (Caspi et al. 2008), MMD (Brown et al. 2009), METLIN (Sana et al. 2008), MZedDB (Draper et al. 2009) and HMDB (Wishart et al. 2009)]. Some databases provide additional information about known metabolic reactions or physiological concentrations (e.g., KEGG and HMDB).

For other identification processes, experimental reference spectra (derived from authentic chemical standards) are required. These are provided by only a limited number of metabolite databases, but are also present in a number of mass spectral libraries. These are observed more frequently in relation to GC–MS for which a number of mass spectral libraries are available, including non-specific (e.g., NIST08; http://chemdata.nist.gov/mass-spc/ms-search/) and metabolite-specific libraries (e.g., GMD; Kopka et al. 2005) and FiehnLib (Kind et al. 2009) and are at present some of the most widely applied mass spectral libraries because of their metabolite coverage. However, a greater number of mass spectral libraries are being constructed. It is important for these libraries that a clear metabolite ontology (Sansone et al. 2007) is used so that valid recognizable identifications are made that are exchangeable between different laboratories.

A number of publicly available databases also contain MS/MS mass spectra. HMDB contains extensive information, mostly about human metabolites, in 8,552 so called MetaboCards. 840 of the compounds have been measured on (mostly low mass resolution) mass spectrometers, and for 916 compounds ^{1}H-NMR spectra are available (Wishart et al. 2009). Unlike other compound databases, the HMDB also contains information about the typical abundance of metabolites in different biofluids and tissues. The METLIN database maintained at the SCRIPPS Institute contains 44,766 compounds, 4,527 of them with high resolution MS/MS spectra (Sana et al. 2008). Recently, many search functions have been added to METLIN, such as an automated batch search: once an mzXML file with MS/MS spectra is uploaded, each MS/MS spectrum can be identified against the METLIN reference spectra. To spread the workload of constructing reference mass spectra in libraries and databases, a consortium approach can be applied. One such example is MassBank (Horai et al. 2010), which operates a number of federated MassBank servers, where a search on one of the nodes will query all other servers in the background, and present the consolidated results. In contrast to the HMDB and METLIN, MassBank has more than 20 different research groups contributing spectra obtained with multiple fragmentation methods, and many of the spectra are licensed under a Creative Commons license (Horai et al. 2010).

6 Current limitations and future outlook

In traditional analytical chemistry, structural elucidation and identification of chemicals is successfully performed for a pure chemical or a simple mixture of chemicals. Mass spectrometry is one of a number of instruments applied in the process of structure elucidation and offers many traditional tools to apply including accurate measurements of *m/z* and acquisition of fragmentation data. The application of MS platforms in untargeted metabolomic studies enables the detection of hundreds or thousands of unique metabolites in a single sample. Applying the traditional MS tools to metabolite identification in simple chemical mixtures is routine. However, their application to the significantly more complex samples studied in untargeted metabolomics is not routine and several limitations and sources for errors are present. Surprisingly, limited assessment of the applicability of these traditional tools to derive metabolite identities in complex samples and their associated accuracy has been performed. The assessment and validation of traditional tools and the development of new tools is an essential requirement for untargeted metabolomic studies to be successful by providing metabolite identification and the ability to derive biological knowledge from metabolomic datasets.

This review has highlighted (i) experimental and computational tools which are currently available and routinely

applied to identify metabolites in mass spectrometry-focused untargeted metabolomic studies, and (ii) new methods that are being developed to increase the accuracy and efficiency of metabolite identification. Over the previous decade significant innovations and developments have been observed. However, we are still at a stage where metabolite identification is a significant bottleneck. Last century it was typical that the identification of ca. 50 % of metabolites was possible in a GC–EI–MS run. Today the number of metabolite features has increased due to enhanced mass spectrometry (shift from quadrupole to ToF separations) and increased mass resolution, but the proportion of identified metabolites has unfortunately stayed the same.

The comparison of experimental data (accurate *m/z*, RT/index, fragmentation mass spectrum) for each metabolite to mass spectral libraries constructed with authentic chemical standards is the ideal process to provide definitive (level 1) identification. This is currently more successfully achieved for data acquired on GC–MS platforms compared to LC–MS or CE–MS platforms. However, mass spectral libraries are limited by the fact that they do not contain all metabolites, and changes in analytical methods or instruments (especially for LC–MS platforms) can render them inaccurate.

Other tools allow reduction of the metabolite search space to a single or small number of metabolites to achieve putative (level 2 or 3) annotation. Further targeted studies can then be performed to confirm identities. These include the collation of data for unidentified metabolites (e.g., MSTs), accurate measurements of *m/z*, acquisition of fragmentation mass spectra related to chemical structure, the application of chemical and biological knowledge (for example, knowledge of experimentally feasible ion formation), experimental isotope-based studies and the development of *in silico* tools to predict mass spectral, chromatographic and electrophoretic properties. This review has documented the high level of innovation in the metabolomics community directed towards developing novel and user-friendly tools for this purpose.

However, further developments and integration of tools are required. Many of the tools have been developed in different research groups (sometimes very similar tools have been developed in multiple laboratories). There have been limited discussions, so far, on the integration of different tools and many laboratories operate a set of separate computational tools rather than an integrated single tool. Limited systematic comparative evaluation of the alternatives is observed in the metabolomics community. The proteomics community has been more pro-active in that respect (e.g. Hoekman et al. 2012). One further improvement required to increase the efficiency of metabolite identification is the integration of different tools, either those focusing on mass spectrometry alone or even aiming at the integration of data from different analytical platforms [e.g., MS and NMR spectroscopy (Crockford et al. 2008)].

One limitation is the lack of metabolites present in mass spectral libraries applied for matching to experimentally derived chromatographic and mass spectral data. It is unrealistic for all metabolites to be purchased or synthesized to allow data to be acquired on authentic and chemically pure metabolites and incorporated into mass spectral libraries. Even when these data are available, the transferability of libraries between instruments can be limited and the development of laboratory-specific mass spectral libraries is costly, ineffective and improbable. However, the development of *in silico* tools to predict mass fragmentation patterns and RTs/indices will provide increasing volumes of data to be incorporated into mass spectral libraries to reduce the number of potential metabolite identifications. However, even if all metabolites were present in mass spectral libraries our current mass spectrometry platforms do not allow the on-line acquisition of MS/MS or MS^n data for all metabolites present in complex metabolomes. Advances in the number of metabolites for which these data are acquired is essential, either applying MS/MS data acquisition on-line with full-scan profiling data or by applying off-line or in-line systems. For any specific analytical method on a single platform, once a metabolite is identified and catalogued (for example, in the form of a MST) it does not need to be identified again as its identity is already known. The identified metabolite can then be applied in mass spectral library searches.

From a biological perspective, one limitation is the size and complexity of specific metabolomes. The plant metabolome (in total across all plants) is estimated to contain more than 200,000 metabolites (Fiehn 2001), of which most are endogenous (although the complexity for a single plant sample will be much lower). Human-derived metabolomes are also complex, as they contain endogenous metabolites in addition to exogenous metabolites acquired from the external environment. The accurate cataloguing of organism-, tissue- or cell-specific metabolites is not yet complete, and significant experimental and informatics resources are required to pursue this cataloguing further. This includes appropriate use of unique identifiers (e.g., InCHI key, SMILES, ChEBI identifier) to allow integration of data from different functional levels in an automated manner in pathway analysis software. Great advances have been made in this direction (for example, within the HMDB project). However, further detailed information is still required. For example, information on drug metabolism is available as text in DrugBank (Wishart et al. 2006), but having these data available as chemical entity identifiers and easily searchable electronically (as is possible in

HMDB for each Metabocard) would be highly advantageous. Not until all metabolites are catalogued electronically, and their physicochemical properties are searchable, can accurate and robust metabolite identification be performed. A metabolite can only be confidently assigned to a metabolic feature if its identity and potential presence is known and reported in databases or mass spectral libraries. Currently this is not possible, and this requires more efficient methods for the de novo structure elucidation of metabolites which are not present in databases.

We are on an important journey to develop the multitude of tools necessary to provide automated and accurate identification of metabolites in complex metabolomic samples. Without the identification of metabolites it is impossible to base biological reasoning on the datasets. We have progressed significantly in recent years, but further developments are essential. In our view, tools to automatically provide definitive (level 1) identification of *all* metabolites in a single sample will not be developed in the near future, but workflows to provide increasingly narrow sets of putative (level 2 or 3) annotations will be improved and combined with subsequent targeted methods for definitive identification. The slow cataloguing of mass spectral data and providing availability to all will increase our knowledge of sample-specific metabolomes. However, the complexity and diversity of metabolomes currently investigated are too limiting in allowing true and complete identification of all metabolites in an automated manner. A community effort is required, hopefully through efforts focused from The Metabolomics Society, to develop the tools and databases and provide integration of these different tools and databases.

Acknowledgments WD and MB gratefully acknowledge support from the National Institute for Health Research (NIHR) Manchester Biomedical Research Centre and the UK NorthWest Development Agency (NWDA). RW thanks both the British Heart Foundation (PG/10/036/28341) and UK Engineering and Physical Sciences Research Council (EP/J501414/1) for support. RG is very grateful to the UK BBSRC for financial support. DJC is funded by an Australian National Health and Medical Research Council (NHMRC) Training Fellowship.

References

An, Z., Chen, Y., Zhang, R., Song, Y., Sun, J., He, J., et al. (2010). Integrated ionization approach for RRLC-MS/MS-based metabonomics: Finding potential biomarkers for lung cancer. *Journal of Proteome Research, 9*(8), 4071–4081.

Beckmann, M., Parker, D., Enot, D. P., Duval, E., & Draper, J. (2008). High-throughput, nontargeted metabolite fingerprinting using nominal mass flow injection electrospray mass spectrometry. *Nature Protocols, 3*(3), 486–504.

Bird, S. S., Marur, V. R., Sniatynski, M. J., Greenberg, H. K., & Kristal, B. S. (2011). Serum lipidomics profiling using LC-MS and high-energy collisional dissociation fragmentation: Focus on triglyceride detection and characterization. *Analytical Chemistry, 83*(17), 6648–6657.

Birkemeyer, C., Kolasa, A., & Kopka, J. (2003). Comprehensive chemical derivatization for gas chromatography-mass spectrometry-based multi-targeted profiling of the major phytohormones. *Journal of Chromatography A, 993*(1–2), 89–102.

Birkemeyer, C., Luedemann, A., Wagner, C., Erban, A., & Kopka, J. (2005). Metabolome analysis: The potential of in vivo labeling with stable isotopes for metabolite profiling. *Trends in Biotechnology, 23*(1), 28–33.

Boroujerdi, A. F., Vizcaino, M. I., Meyers, A., Pollock, E. C., Huynh, S. L., Schock, T. B., et al. (2009). NMR-based microbial metabolomics and the temperature-dependent coral pathogen *Vibrio coralliilyticus*. *Environmental Science and Technology, 43*(20), 7658–7664.

Breitling, R., Pitt, A. R., & Barrett, M. P. (2006a). Precision mapping of the metabolome. *Trends in Biotechnology, 24*(12), 543–548.

Breitling, R., Ritchie, S., Goodenowe, D., Stewart, M. L., & Barrett, M. P. (2006b). Ab initio prediction of metabolic networks using Fourier transform mass spectrometry data. *Metabolomics, 2*(3), 155–164.

Brown, M., Dunn, W. B., Dobson, P., Patel, Y., Winder, C. L., Francis-McIntyre, S., et al. (2009). Mass spectrometry tools and metabolite-specific databases for molecular identification in metabolomics. *Analyst, 134*(7), 1322–1332.

Brown, M., Wedge, D. C., Goodacre, R., Kell, D. B., Baker, P. N., Kenny, L. C., et al. (2011). Automated workflows for accurate mass-based putative metabolite identification in LC/MS-derived metabolomic datasets. *Bioinformatics, 27*(8), 1108–1112.

Caspi, R., Foerster, H., Fulcher, C. A., Kaipa, P., Krummenacker, M., Latendresse, M., et al. (2008). The MetaCyc Database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases. *Nucleic Acids Research, 36*(Database issue), D623–D631.

Castro-Perez, J., Roddy, T. P., Nibbering, N. M., Shah, V., McLaren, D. G., Previs, S., et al. (2011). Localization of fatty acyl and double bond positions in phosphatidylcholines using a dual stage CID fragmentation coupled with ion mobility mass spectrometry. *Journal of the American Society for Mass Spectrometry, 22*(9), 1552–1567.

Creek, D. J., Jankevics, A., Breitling, R., Watson, D. G., Barrett, M. P., & Burgess, K. E. (2011). Toward global metabolomics analysis with hydrophilic interaction liquid chromatography-mass spectrometry: Improved metabolite identification by retention time prediction. *Analytical Chemistry, 83*(22), 8703–8710.

Creek, D. J., Jankevics, A., Burgess, K. E., Breitling, R., & Barrett, M. P. (2012). IDEOM: An Excel interface for analysis of LC-MS based metabolomics data. *Bioinformatics, 28*(7), 1048–1049.

Crockford, D. J., Maher, A. D., Ahmadi, K. R., Barrett, A., Plumb, R. S., Wilson, I. D., et al. (2008). 1H NMR and UPLC-MS(E) statistical heterospectroscopy: Characterization of drug metabolites (xenometabolome) in epidemiological studies." *Analytical Chemistry, 80*(18), 6835–6844.

de Hoffmann, E., & Stroobant, V. (2007). *Mass spectrometry—Principle and applications*. Chichester: Wiley.

De Vos, R. C., Moco, S., Lommen, A., Keurentjes, J. J., Bino, R. J., & Hall, R. D. (2007). Untargeted large-scale plant metabolomics using liquid chromatography coupled to mass spectrometry. *Nature Protocols, 2*(4), 778–791.

Dear, G. J., Plumb, R. S., Sweatman, B. C., Ismail, I. M., & Ayrton, J. (1999). Tandem mass spectrometry linked fraction collection for the isolation of drug metabolites from biological matrices. *Rapid Communications in Mass Spectrometry, 13*(10), 886–894.

Desbrosses, G. G., Kopka, J., & Udvardi, M. K. (2005). *Lotus japonicus* metabolic profiling. Development of gas chromatography-mass

spectrometry resources for the study of plant–microbe interactions. *Plant Physiology, 137*(4), 1302–1318.

Dettmer, K., Aronov, P. A., & Hammock, B. D. (2007). Mass spectrometry-based metabolomics. *Mass Spectrometry Reviews, 26*(1), 51–78.

Draper, J., Enot, D. P., Parker, D., Beckmann, M., Snowdon, S., Lin, W., et al. (2009). Metabolite signal identification in accurate mass metabolomics data with MZedDB, an interactive m/z annotation tool utilising predicted ionisation behaviour 'rules'. *BMC Bioinformatics, 10*(1), 227.

Dunn, W. B. (2008). Current trends and future requirements for the mass spectrometric investigation of microbial, mammalian and plant metabolomes. *Physical Biology, 5*(1), 011001.

Dunn, W. B., Broadhurst, D. I., Atherton, H. J., Goodacre, R., & Griffin, J. L. (2011a). Systems level studies of mammalian metabolomes: The roles of mass spectrometry and nuclear magnetic resonance spectroscopy. *Chemical Society Reviews, 40*(1), 387–426.

Dunn, W. B., Broadhurst, D., Begley, P., Zelena, E., Francis-McIntyre, S., Anderson, N., et al. (2011b). Procedures for large-scale metabolic profiling of serum and plasma using gas chromatography and liquid chromatography coupled to mass spectrometry. *Nature Protocols, 6*(7), 1060–1083.

Dunn, W. B., Brown, M., Worton, S. A., Crocker, I. P., Broadhurst, D., Horgan, R., et al. (2009). Changes in the metabolic footprint of placental explant-conditioned culture medium identifies metabolic disturbances related to hypoxia and pre-eclampsia. *Placenta, 30*(11), 974–980.

Erve, J. C., Gu, M., Wang, Y., DeMaio, W., & Talaat, R. E. (2009). Spectral accuracy of molecular ions in an LTQ/Orbitrap mass spectrometer and implications for elemental composition determination. *Journal of the American Society for Mass Spectrometry, 20*(11), 2058–2069.

Eyres, G. T., Urban, S., Morrison, P. D., Dufour, J. P., & Marriott, P. J. (2008). Method for small-molecule discovery based on microscale-preparative multidimensional gas chromatography isolation with nuclear magnetic resonance spectroscopy. *Analytical Chemistry, 80*(16), 6293–6299.

Farag, M. A., Deavours, B. E., de Fátima, A., Naoumkina, M., Dixon, R. A., & Sumner, L. W. (2009). Integrated metabolite and transcript profiling identify a biosynthetic mechanism for hispidol in *Medicago truncatula* cell cultures. *Plant Physiology, 151*(3), 1096–1113.

Feldberg, L., Venger, I., Malitsky, S., Rogachev, I., & Aharoni, A. (2009). Dual labeling of metabolites for metabolome analysis (DLEMMA): A new approach for the identification and relative quantification of metabolites by means of dual isotope labeling and liquid chromatography-mass spectrometry. *Analytical Chemistry, 81*(22), 9257–9266.

Fenn, J. B., Mann, M., Meng, C. K., Wong, S. F., & Whitehouse, C. M. (1989). Electrospray ionization for mass spectrometry of large biomolecules. *Science, 246*(4926), 64–71.

Fernie, A. R., Aharoni, A., Willmitzer, L., Stitt, M., Tohge, T., Kopka, J., et al. (2011). Recommendations for reporting metabolite data. *The Plant Cell, 23*(7), 2477–2482.

Fiehn, O. (2001). Combining genomics, metabolome analysis, and biochemical modelling to understand metabolic networks. *Comparative and Functional Genomics, 2*(3), 155–168.

Fiehn, O. (2002). Metabolomics—The link between genotypes and phenotypes. *Plant Molecular Biology, 48*(1–2), 155–171.

Fiehn, O., Kopka, J., Dörmann, P., Altmann, T., Trethewey, R. N., & Willmitzer, L. (2000a). Metabolite profiling for plant functional genomics. *Nature Biotechnology, 18*(11), 1157–1161.

Fiehn, O., Kopka, J., Trethewey, R. N., & Willmitzer, L. (2000b). Identification of uncommon plant metabolites based on calculation of elemental compositions using gas chromatography and quadrupole mass spectrometry. *Analytical Chemistry, 72*(15), 3573–3580.

Fuhrer, T., Heer, D., Begemann, B., & Zamboni, N. (2011). High-throughput, accurate mass metabolome profiling of cellular extracts by flow injection-time-of-flight mass spectrometry. *Analytical Chemistry, 83*(18), 7074–7080.

Giavalisco, P., Hummel, J., Lisec, J., Inostroza, A. C., Catchpole, G., & Willmitzer, L. (2008). High-resolution direct infusion-based mass spectrometry in combination with whole 13C metabolome isotope labeling allows unambiguous assignment of chemical sum formulas. *Analytical Chemistry, 80*(24), 9417–9425.

Giavalisco, P., Köhl, K., Hummel, J., Seiwert, B., & Willmitzer, L. (2009). 13C isotope-labeled metabolomes allowing for improved compound annotation and relative quantification in liquid chromatography-mass spectrometry-based metabolomic research. *Analytical Chemistry, 81*(15), 6546–6551.

Gipson, G. T., Tatsuoka, K. S., Sokhansanj, B. A., Ball, R. J., & Connor, S. C. (2008). Assignment of MS-based metabolomic datasets via compound interaction pair mapping. *Metabolomics, 4*(1), 94–103.

Goodacre, R. (2007). Metabolomics of a superorganism. *Journal of Nutrition, 137*(1 Suppl), 259S–266S.

Halket, J. M., & Zaikin, V. G. (2003). Derivatization in mass spectrometry—1. Silylation. *European Journal of Mass Spectrometry, 9*(1), 1–21.

Heinonen, M., Rantanen, A., Mielikäinen, T., Kokkonen, J., Kiuru, J., Ketola, R. A., et al. (2008). FiD: A software for ab initio structural identification of product ions from tandem mass spectrometric data. *Rapid Communications in Mass Spectrometry, 22*(19), 3043–3052.

Herrgård, M. J., Swainston, N., Dobson, P., Dunn, W. B., Arga, K. Y., Arvas, M., et al. (2008). A consensus yeast metabolic network reconstruction obtained from a community approach to systems biology. *Nature Biotechnology, 26*(10), 1155–1160.

Hildebrandt, C., Wolf, S., & Neumann, S. (2011). Database supported candidate search for metabolite identification. *Journal of Integrative Bioinformatics, 8*(2), 157.

Hill, D. W., Kertesz, T. M., Fontaine, D., Friedman, R., & Grant, D. F. (2008). Mass spectral metabonomics beyond elemental formula: Chemical database querying by matching experimental with computational fragmentation spectra. *Analytical Chemistry, 80*(14), 5574–5582.

Hill, A. W., & Mortishire-Smith, R. J. (2005). Automated assignment of high-resolution collisionally activated dissociation mass spectra using a systematic bond disconnection approach. *Rapid Communications in Mass Spectrometry, 19*(21), 3111–3118.

Hoekman, B., Breitling, R., Suits, F., Bischoff, R., & Horvatovich, P. (2012). msCompare: A framework for quantitative analysis of label-free LC-MS data for comparative biomarker studies. *Molecular & Cellular Proteomics*. doi:10.1074/mcp.M111.015974.

Hoopmann, M. R., Merrihew, G. E., von Haller, P. D., & MacCoss, M. J. (2009). Post analysis data acquisition for the iterative MS/MS sampling of proteomics mixtures. *Journal of Proteome Research, 8*(4), 1870–1875.

Horai, H., Arita, M., Kanaya, S., Nihei, Y., Ikeda, T., Suwa, K., et al. (2010). MassBank: A public repository for sharing mass spectral data for life sciences. *Journal of Mass Spectrometry, 45*(7), 703–714.

Huege, J., Goetze, J., Schwarz, D., Bauwe, H., Hagemann, M., & Kopka, J. (2011). Modulation of the major paths of carbon in photorespiratory mutants of synechocystis. *PLoS ONE, 6*(1), e16278.

Huege, J., Sulpice, R., Gibon, Y., Lisec, J., Koehl, K., & Kopka, J. (2007). GC-EI-TOF-MS analysis of in vivo carbon-partitioning into soluble metabolite pools of higher plants by monitoring

isotope dilution after $^{13}CO_2$ labelling. *Phytochemistry, 68*(16–18), 2258–2272.

Hummel, J., Strehmel, N., Selbig, J., Walther, D., & Kopka, J. (2010). Decision tree supported substructure prediction of metabolites from GC-MS profiles. *Metabolomics, 6*(2), 322–333.

Iijima, Y., Nakamura, Y., Ogata, Y., Tanaka, K., Sakurai, N., Suda, K., et al. (2008). Metabolite annotations based on the integration of mass spectral information. *The Plant Journal, 54*(5), 949–962.

Kahar, P., Taku, K., & Tanaka, S. (2011). Enhancement of xylose uptake in 2-deoxyglucose tolerant mutant of *Saccharomyces cerevisiae*. *Journal of Bioscience and Bioengineering, 111*(5), 557–563.

Kanehisa, M., Goto, S., Furumichi, M., Tanabe, M., & Hirakawa, M. (2010). KEGG for representation and analysis of molecular networks involving diseases and drugs. *Nucleic Acids Research, 38*(Database issue), D355–D360.

Kaufmann, A. (2010). Strategy for the elucidation of elemental compositions of trace analytes based on a mass resolution of 100,000 full width at half maximum. *Rapid Communications in Mass Spectrometry, 24*(14), 2035–2045.

Kenny, L. C., Broadhurst, D. I., Dunn, W., Brown, M., North, R. A., McCowan, L., et al. (2010). Robust early pregnancy prediction of later preeclampsia using metabolomic biomarkers. *Hypertension, 56*(4), 741–749.

Kind, T., & Fiehn, O. (2006). Metabolomic database annotations via query of elemental compositions: Mass accuracy is insufficient even at less than 1 ppm. *BMC Bioinformatics, 7*, 234.

Kind, T., & Fiehn, O. (2007). Seven Golden Rules for heuristic filtering of molecular formulas obtained by accurate mass spectrometry. *BMC Bioinformatics, 8*, 105.

Kind, T., & Fiehn, O. (2011). Advances in structure elucidation of small molecules using mass spectrometry. *Bioanalytical Reviews, 2*(1–4), 23–60.

Kind, T., Wohlgemuth, G., Lee, D. Y., Lu, Y., Palazoglu, M., Shahbaz, S., et al. (2009). FiehnLib: Mass spectral and retention index libraries for metabolomics based on quadrupole and time-of-flight gas chromatography/mass spectrometry. *Analytical Chemistry, 81*(24), 10038–10048.

Kirchmair, J., Williamson, M. J., Tyzack, J. D., Tan, L., Bond, P. J., Bender, A., et al. (2012). Computational prediction of metabolism: Sites, products, SAR, P450 enzyme dynamics, and mechanisms. *Journal of Chemical Information and Modeling, 52*(3), 617–648.

Koch, B. P., Dittmar, T., Witt, M., & Kattner, G. (2007). Fundamentals of molecular formula assignment to ultrahigh resolution mass data of natural organic matter. *Analytical Chemistry, 79*(4), 1758–1763.

Komatsu, M., Uchiyama, T., Omura, S., Cane, D. E., & Ikeda, H. (2010). Genome-minimized Streptomyces host for the heterologous expression of secondary metabolism. *The Proceedings of the National Academy of Sciences of the United States of America, 107*(6), 2646–2651.

Konishi, Y., Kiyota, T., Draghici, C., Gao, J. M., Yeboah, F., Acoca, S., et al. (2007). Molecular formula analysis by an MS/MS/MS technique to expedite dereplication of natural products. *Analytical Chemistry, 79*(3), 1187–1197.

Kopka, J. (2006). Current challenges and developments in GC-MS based metabolite profiling technology. *Journal of Biotechnology, 124*(1), 312–322.

Kopka, J., Schauer, N., Krueger, S., Birkemeyer, C., Usadel, B., Bergmüller, E., et al. (2005). GMD@CSB.DB: The Golm Metabolome Database. *Bioinformatics, 21*(8), 1635–1638.

Kuhl, C., Tautenhahn, R., Böttcher, C., Larson, T. R., & Neumann, S. (2011). CAMERA: An integrated strategy for compound spectra extraction and annotation of liquid chromatography/mass spectrometry data sets. *Analytical Chemistry, 84*(1), 283–289.

Kumari, S., Stevens, D., Kind, T., Denkert, C., & Fiehn, O. (2011). Applying in silico retention index and mass spectra matching for identification of unknown metabolites in accurate mass GC-TOF mass spectrometry. *Analytical Chemistry, 83*(15), 5895–5902.

Lei, Z., Huhman, D. V., & Sumner, L. W. (2011). Mass spectrometry strategies in metabolomics. *Journal of Biological Chemistry, 286*(29), 25435–25442.

Lisec, J., Schauer, N., Kopka, J., Willmitzer, L., & Fernie, A. R. (2006). Gas chromatography mass spectrometry-based metabolite profiling in plants. *Nature Protocols, 1*(1), 387–396.

Lloyd, A. J., Beckmann, M., Favé, G., Mathers, J. C., & Draper, J. (2011). Proline betaine and its biotransformation products in fasting urine samples are potential biomarkers of habitual citrus fruit consumption. *British Journal of Nutrition, 106*(6), 812–824.

Loo, R. L., Chan, Q., Brown, I. J., Robertson, C. E., Stamler, J., Nicholson, J. K., et al. (2012). A comparison of self-reported analgesic use and detection of urinary ibuprofen and acetaminophen metabolites by means of metabonomics: The INTERMAP study. *American Journal of Epidemiology, 175*(4), 348–358.

Lu, X., Zhao, X., Bai, C., Zhao, C., Lu, G., & Xu, G. (2008). LC-MS-based metabonomics analysis. *Journal of Chromatography B—Analytical Technologies in the Biomedical and Life Sciences, 866*(1–2), 64–76.

Lugan, R., Niogret, M. F., Leport, L., Guégan, J. P., Larher, F. R., Savouré, A., et al. (2010). Metabolome and water homeostasis analysis of *Thellungiella salsuginea* suggests that dehydration tolerance is a key response to osmotic stress in this halophyte. *The Plant Journal, 64*(2), 215–229.

Malvoisin, E., Evrard, E., Roberfroid, M., & Mercier, M. (1979). Determination of Kovats retention indices with a capillary column and electron-capture detection: Application to the assay of the enzymatic conversion of 3,4-epoxy-1-butene into diepoxybutane. *Journal of Chromatography, 186*, 81–87.

Matsuda, F., Shinbo, Y., Oikawa, A., Hirai, M. Y., Fiehn, O., Kanaya, S., et al. (2009). Assessment of metabolome annotation quality: A method for evaluating the false discovery rate of elemental composition searches. *PLoS ONE, 4*(10), e7490.

Mihaleva, V. V., Verhoeven, H. A., de Vos, R. C., Hall, R. D., & van Ham, R. C. (2009). Automated procedure for candidate compound selection in GC-MS metabolomics based on prediction of Kovats retention index. *Bioinformatics, 25*(6), 787–794.

Miura, D., Tsuji, Y., Takahashi, K., Wariishi, H., & Saito, K. (2010). A strategy for the determination of the elemental composition by Fourier transform ion cyclotron resonance mass spectrometry based on isotopic peak ratios. *Analytical Chemistry, 82*(13), 5887–5891.

Neumann, S., Thum, A., & Böttcher, S. (2012). Nearline acquisition and processing of liquid chromatography-tandem mass spectrometry data. *Metabolomics*. doi:10.1007/s11306-012-0401-0.

Ochiai, N., & Sasamoto, K. (2010). Selectable one-dimensional or two-dimensional gas chromatography-olfactometry/mass spectrometry with preparative fraction collection for analysis of ultra-trace amounts of odor compounds. *Journal of Chromatography A, 1218*(21), 3180–3185.

Ogata, H., Goto, S., Sato, K., Fujibuchi, W., Bono, H., & Kanehisa, M. (1999). KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Research, 27*(1), 29–34.

Oresic, M., Simell, S., Sysi-Aho, M., Näntö-Salonen, K., Seppänen-Laakso, T., Parikka, V., et al. (2008). Dysregulation of lipid and amino acid metabolism precedes islet autoimmunity in children who later progress to type 1 diabetes. *Journal of Experimental Medicine, 205*(13), 2975–2984.

Pechlivanis, A., Kostidis, S., Saraslanidis, P., Petridou, A., Tsalis, G., Mougios, V., et al. (2010). (1)H-NMR-based metabonomic investigation of the effect of two different exercise sessions on

the metabolic fingerprint of human urine. *Journal of Proteome Research, 9*(12), 6405–6416.

Plumb, R. S., Johnson, K. A., Rainville, P., Smith, B. W., Wilson, I. D., Castro-Perez, J. M., et al. (2006). UPLIC/MS(E); A new approach for generating molecular fragment information for biomarker structure elucidation. *Rapid Communications in Mass Spectrometry, 20*(13), 1989–1994.

Pope, G. A., MacKenzie, D. A., Defernez, M., Aroso, M. A., Fuller, L. J., Mellon, F. A., et al. (2007). Metabolic footprinting as a tool for discriminating between brewing yeasts. *Yeast, 24*(8), 667–679.

Ramautar, R., Somsen, G. W., & de Jong, G. J. (2009). CE-MS in metabolomics. *Electrophoresis, 30*(1), 276–291.

Roessner, U., Luedemann, A., Brust, D., Fiehn, O., Linke, T., Willmitzer, L., et al. (2001). Metabolic profiling allows comprehensive phenotyping of genetically or environmentally modified plant systems. *Plant Cell, 13*(1), 11–29.

Roessner, U., Wagner, C., Kopka, J., Trethewey, R. N., & Willmitzer, L. (2000). Technical advance: Simultaneous analysis of metabolites in potato tuber by gas chromatography-mass spectrometry. *The Plant Journal, 23*(1), 131–142.

Rogers, S., Scheltema, R. A., Girolami, M., & Breitling, R. (2009). Probabilistic assignment of formulas to mass peaks in metabolomics experiments. *Bioinformatics, 25*(4), 512–518.

Rojas-Chertó, M., Kasper, P. T., Willighagen, E. L., Vreeken, R. J., Hankemeier, T., & Reijmers, T. H. (2011). Elemental composition determination based on MS(n). *Bioinformatics, 27*(17), 2376–2383.

Sana, T. R., Roark, J. C., Li, X., Waddell, K., & Fischer, S. M. (2008). Molecular formula and METLIN Personal Metabolite Database matching applied to the identification of compounds generated by LC/TOF-MS. *Journal of Biomolecular Techniques, 19*(4), 258–266.

Sansone, S.-A., Schober, D., Atherton, H. J., Fiehn, O., Jenkins, H., Rocca-Serra, P., et al. (2007). Metabolomics standards initiative—Ontology working group work in progress. *Metabolomics, 3*(3), 249–256.

Schauer, N., Steinhauser, D., Strelkov, S., Schomburg, D., Allison, G., Moritz, T., et al. (2005). GC-MS libraries for the rapid identification of metabolites in complex biological samples. *FEBS Letters, 579*(6), 1332–1337.

Scheltema, R. A., Jankevics, A., Jansen, R. C., Swertz, M. A., & Breitling, R. (2011). PeakML/mzMatch: A file format, Java library, R library, and tool-chain for mass spectrometry data analysis. *Analytical Chemistry, 83*(7), 2786–2793.

Scheltema, R. A., Kamleh, A., Wildridge, D., Ebikeme, C., Watson, D. G., Barrett, M. P., et al. (2008). Increasing the mass accuracy of high-resolution LC-MS data using background ions: A case study on the LTQ-Orbitrap. *Proteomics, 8*(22), 4647–4656.

Schmidt, B., Joussen, N., Bode, M., & Schuphan, I. (2006). Oxidative metabolic profiling of xenobiotics by human P450s expressed in tobacco cell suspension cultures. *Biochemical Society Transactions, 34*(Pt 6), 1241–1245.

Schug, K., & McNair, H. M. (2002). Adduct formation in electrospray ionization. Part 1: Common acidic pharmaceuticals. *Journal of Separation Science, 25*(12), 759–766.

Schug, K., & McNair, H. M. (2003). Adduct formation in electrospray ionization mass spectrometry II. Benzoic acid derivatives. *Journal of Chromatography A, 985*(1–2), 531–539.

Schymanski, E. L., Gallampois, C. M., Krauss, M., Meringer, M., Neumann, S., Schulze, T., et al. (2012). Consensus structure elucidation combining GC/EI-MS, structure generation and calculated properties. *Analytical Chemistry, 84*(7), 3287–3295.

Schymanski, E. L., Meringer, M., & Brack, W. (2011). Automated strategies to identify compounds on the basis of GC/EI-MS and calculated properties. *Analytical Chemistry, 83*(3), 903–912.

Sheldon, M. T., Mistrik, R., & Croley, T. R. (2009). Determination of ion structures in structurally related compounds using precursor ion fingerprinting. *Journal of the American Society for Mass Spectrometry, 20*(3), 370–376.

Siegel, M. M., & Gill, G. (1990). MASSPEC: A graphics-based data system for correlating a mass spectrum with a proposed structure. *Analytica Chimica Acta, 237*, 459–472.

Smart, K. F., Aggio, R. B., Van Houtte, J. R., & Villas-Bôas, S. G. (2010). Analytical platform for metabolome analysis of microbial cells using methyl chloroformate derivatization followed by gas chromatography-mass spectrometry. *Nature Protocols, 5*(10), 1709–1729.

Soga, T., Ohashi, Y., Ueno, Y., Naraoka, H., Tomita, M., & Nishioka, T. (2003). Quantitative metabolome analysis using capillary electrophoresis mass spectrometry. *Journal of Proteome Research, 2*(5), 488–494.

Southam, A. D., Payne, T. G., Cooper, H. J., Arvanitis, T. N., & Viant, M. R. (2007). Dynamic range and mass accuracy of wide-scan direct infusion nanoelectrospray Fourier transform ion cyclotron resonance mass spectrometry-based metabolomics increased by the spectral stitching method. *Analytical Chemistry, 79*(12), 4595–4602.

Spagou, K., Wilson, I. D., Masson, P., Theodoridis, G., Raikos, N., Coen, M., et al. (2010). HILIC-UPLC-MS for exploratory urinary metabolic profiling in toxicological studies. *Analytical Chemistry, 83*(1), 382–390.

Stoll, N., Schmidt, E., & Thurow, K. (2006). Isotope pattern evaluation for the reduction of elemental compositions assigned to high-resolution mass spectral data from electrospray ionization Fourier transform ion cyclotron resonance mass spectrometry. *Journal of the American Society for Mass Spectrometry, 17*(12), 1692–1699.

Strehmel, N., Hummel, J., Erban, A., Strassburg, K., & Kopka, J. (2008). Retention index thresholds for compound matching in GC-MS metabolite profiling. *The Journal of Chromatography B: Analytical Technologies in the Biomedical and Life Sciences, 871*(2), 182–190.

Sud, M., Fahy, E., Cotter, D., Brown, A., Dennis, E. A., Glass, C. K., et al. (2007). LMSD: LIPID MAPS structure database. *Nucleic Acids Res, 35*(Database issue), D527–D532.

Sugimoto, M., Kikuchi, S., Arita, M., Soga, T., Nishioka, T., & Tomita, M. (2005). Large-scale prediction of cationic metabolite identity and migration time in capillary electrophoresis mass spectrometry using artificial neural networks. *Analytical Chemistry, 77*(1), 78–84.

Sumner, L. W., Amberg, A., Barrett, D., Beale, M. H., Beger, R., Daykin, C. A., et al. (2007). Proposed minimum reporting standards for chemical analysis. *Metabolomics, 3*(3), 211–221.

Taylor, N. S., Weber, R. J. M., Southam, A. D., Payne, T. G., Hrydziuszko, O., Arvanitis, T. N., et al. (2009). A new approach to toxicity testing in *Daphnia magna*: Application of high throughput FT-ICR mass spectrometry metabolomics. *Metabolomics, 5*(1), 44–58.

Theodoridis, G., Gika, H. G., & Wilson, I. D. (2008). LC-MS-based methodology for global metabolite profiling in metabonomics/metabolomics. *TrAC—Trends in Analytical Chemistry, 27*(3), 251–260.

Tikunov, Y., Lommen, A., de Vos, C. H., Verhoeven, H. A., Bino, R. J., Hall, R. D., et al. (2005). A novel approach for nontargeted data analysis for metabolomics. Large-scale profiling of tomato fruit volatiles. *Plant Physiology, 139*(3), 1125–1137.

Tong, H., Bell, D., Tabei, K., & Siegel, M. M. (1999). Automated data massaging, interpretation, and e-mailing modules for high throughput open access mass spectrometry. *Journal of the American Society for Mass Spectrometry, 10*(11), 1174–1187.

van der Hooft, J. J., Vervoort, J., Bino, R. J., Beekwilder, J., & de Vos, R. C. (2010). Polyphenol identification based on systematic and robust high-resolution accurate mass spectrometry fragmentation. *Analytical Chemistry, 83*(1), 409–416.

van der Werf, M. J., Overkamp, K. M., Muilwijk, B., Coulier, L., & Hankemeier, T. (2007). Microbial metabolomics: Toward a platform with full metabolome coverage. *Analytical Biochemistry, 370*(1), 17–25.

Viant, M. R. (2008). Recent developments in environmental metabolomics. *Molecular BioSystems, 4*(10), 980–986.

Wachsmuth, C. J., Almstetter, M. F., Waldhier, M. C., Gruber, M. A., Nürnberger, N., Oefner, P. J., et al. (2011). Performance evaluation of gas chromatography-atmospheric pressure chemical ionization-time-of-flight mass spectrometry for metabolic fingerprinting and profiling. *Analytical Chemistry, 83*(19), 7514–7522.

Wagner, C., Sefkow, M., & Kopka, J. (2003). Construction and application of a mass spectral and retention time index database generated from plant GC/EI-TOF-MS metabolite profiles. *Phytochemistry, 62*(6), 887–900.

Wang, X., Liang, Y., Zhu, L., Xie, H., Li, H., He, J., et al. (2010). Preparative isolation and purification of flavone c-glycosides from the leaves of *Ficus microcarpa* L. f by medium-pressure liquid chromatography, high-speed countercurrent chromatography, and preparative liquid chromatography. *Journal of Liquid Chromatography & Related Technologies, 33*(4), 462–480.

Weber, R. J., Southam, A. D., Sommer, U., & Viant, M. R. (2011). Characterization of isotopic abundance measurements in high resolution FT-ICR and orbitrap mass spectra for improved confidence of metabolite identification. *Analytical Chemistry, 83*(10), 3737–3743.

Weber, R. J. M., & Viant, M. R. (2010). MI-Pack: Increased confidence of metabolite identification in mass spectra by integrating accurate masses and metabolic pathways. *Chemometrics and Intelligent Laboratory Systems, 104*(1), 75–82.

Welthagen, W., Shellie, R. A., Spranger, J., Ristow, M., Zimmermann, R., & Fiehn, O. (2005). Comprehensive two-dimensional gas chromatography-time-of-flight mass spectrometry (GC × GC-TOF) for high resolution metabolomics: Biomarker discovery on spleen tissue extracts of obese NZO compared to lean C57BL/6 mice. *Metabolomics, 1*(1), 65–73.

Wikoff, W. R., Anfora, A. T., Liu, J., Schultz, P. G., Lesley, S. A., Peters, E. C., et al. (2009). Metabolomics analysis reveals large effects of gut microflora on mammalian blood metabolites. *The Proceedings of the National Academy of Sciences of the United States of America, 106*(10), 3698–3703.

Winder, C. L., Dunn, W. B., & Goodacre, R. (2011). TARDIS-based microbial metabolomics: Time and relative differences in systems. *Trends in Microbiology, 19*(7), 315–322.

Wishart, D. S. (2011). Advances in metabolite identification. *Bioanalysis, 3*(15), 1769–1782.

Wishart, D. S., Knox, C., Guo, A. C., Eisner, R., Young, N., Gautam, B., et al. (2009). HMDB: A knowledgebase for the human metabolome. *Database issue, 37*(3), D603–D610.

Wishart, D. S., Knox, C., Guo, A. C., Shrivastava, S., Hassanali, M., Stothard, P., et al. (2006). DrugBank: A comprehensive resource for in silico drug discovery and exploration. *Nucleic Acids Research, 34*(Database issue), D668–D672.

Wolf, S., Schmidt, S., Müller-Hannemann, M., & Neumann, S. (2010). In silico fragmentation for computer assisted identification of metabolite mass spectra. *BMC Bioinformatics, 11*, 148.

Xu, Y., Heilier, J. F., Madalinski, G., Genin, E., Ezan, E., Tabet, J. C., et al. (2010). Evaluation of accurate mass and relative isotopic abundance measurements in the LTQ-orbitrap mass spectrometer for further metabolomics database building. *Analytical Chemistry, 82*(13), 5490–5501.

Yuan, J., Doucette, C. D., Fowler, W. U., Feng, X. J., Piazza, M., Rabitz, H. A., et al. (2009). Metabolomics-driven quantitative analysis of ammonia assimilation in *E. coli*. *Molecular Systems Biology, 5*, 302.

Zhu, J., & Cole, R. B. (2000). Formation and decompositions of chloride adduct ions. *Journal of the American Society for Mass Spectrometry, 11*(11), 932–941.

Zhudamo, J. S., Qunfa Hong, R. L., Lu, P., & Wang, L. (1988). ASES/MS: An automatic structure elucidation system for organic compounds using mass spectrometric data. *The Analyst, 113*, 1261–1265.